# AUC Undergraduate Journal of Liberal Arts & Sciences

## Capstone Issue Vol. 12 2019

www.auc.nl

University of Amsterdam

VU University Amsterdam

# AUC Undergraduate Journal of Liberal Arts & Sciences

Capstone Issue Vol. 12 2019

Published by
inprint.

**AUC Undergraduate Journal of Liberal Arts & Sciences**
**Capstone Issue Vol. 12 2019, published by InPrint**

The AUC Undergraduate Journal of Liberal Arts and Sciences is a biannual, interdisciplinary publication showcasing outstanding undergraduate academic papers. The Journal aims to demonstrate the strength of undergraduate scholarship at AUC, to reflect the intellectual diversity of its academic programme, to encourage best research and writing practices, to facilitate collaboration between students and faculty across the curriculum, and to provide students with opportunities to gain experience in academic reviewing, editing and publishing. The Editorial of the Journal is constituted of members of the InPrint board, a registered AUCSA committee.

### Editorial board (InPrint)

### Capstone Team

### Series Editing

## Foreword

Before they graduate, AUC students are required to perform an independent research project within their intended majors of Sciences, Social Sciences or the Humanities: the Capstone. In the semester they spend writing their Capstones, students learn to engage and contribute to the academic dialogue occurring within their chosen field. With the current Capstone Issue, we publish six Capstones submitted to us from across AUC's three majors by the graduating students of 2019 which demonstrate AUC students' diverse skills, interests and academic excellence.

All the Capstones published in this issue have undergone a rigorous selection and editing process carried out by our Editorial Board to improve the clarity and accessibility of the selected works. We'd also like to thank Dr. Lotte Tavecchio and Dean Prof. Dr. Murray Pratt for their support and enthusiasm for our publication. Additionally, a thank you is due for the Academic Writing Skills and Advanced Research Writing teams, who through the support and inclusion of our journal in their classes have helped ensure InPrint's longevity and presence within the AUC community. Finally, the template for the $\LaTeX\,2_{\varepsilon}$ formatting for this issue was graciously provided by former Editor-in-Chief Lanie Preston.

I sincerely hope that in reading this Capstone Issue you, our dear reader, can see the high level of academic achievement of AUC students, their incredible work ethic, and the wonderful variety in interests and topics that is emblematic of the liberal arts and sciences education. And now, without further ado, I welcome you to our 12th Capstone Issue!

*Aisha Erenstein, on behalf of InPrint*

# Contents

**Sciences**

**Sciences**

**Social Sciences**

**Social Sciences**

**Humanities - Social Sciences**

**Humanities**

Sciences

# Implementing Quantum-Cryptographic Protocols using SimulaQron

A quantum-internet simulation of 1-2 oblivious transfer in the noisy-storage model

Lynn Engelberts

*Supervisor*
Dr. Christian Schaffner (UvA, QuSoft)
*Reader*
Dr. Michael Walter (UvA, QuSoft)

### Abstract

Research in quantum technology is currently devoted to the development of a large-scale quantum network, also known as a quantum internet. Since a quantum internet is expected to be realised in the near future, a quantum-internet simulator, SimulaQron, has been created to function as a framework for the development of quantum-internet applications [DW18]. One of the expected applications of a quantum internet is the cryptographic primitive secure two-party computation. However, the suitability of SimulaQron for cryptographic applications, such as secure two-party computation, has not yet been examined. In particular, this thesis demonstrates the potential and limitations of SimulaQron as environment for implementing and analysing cryptographic protocols. Furthermore, we evaluate whether the considered protocols are suitable for quantum-internet applications. To the best of our knowledge, this thesis provides the first implementations of quantum-cryptographic protocols in SimulaQron. Our implementations are based on the protocols for 1-2 oblivious transfer (1-2 OT) proposed in [Sch10]. 1-2 OT is a simple two-party primitive on which any protocol for secure two-party computation can be based [GV88; Kil88]. Assuming that the laws of quantum mechanics hold, the protocols in [Sch10] are shown to be secure against attacks in the noisy-storage model – a realistic cryptographic model which assumes that the adversary's quantum storage is noisy. Our work is based on a critical literature analysis of 1-2 OT in the noisy-storage model and quantum internet, followed by an implementation of the protocols provided in [Sch10] using SimulaQron. Our implementations will allow us to evaluate whether these protocols are easy to implement and suitable for quantum internet. This research will thus contribute to both the field of quantum cryptography and the area of quantum-internet development by analysing 1-2 OT in the noisy-storage model as well as evaluating the cryptographic applications of a quantum internet and the potential of SimulaQron.

Keywords and phrases: *1-2 oblivious transfer (1-2 OT), error correction, noisy-storage model, quantum internet, SimulaQron*

## Acknowledgements

Firstly, I would like to thank my supervisor Christian Schaffner for his patience and guidance during this project. By introducing me to the area of quantum cryptography and information theory, Christian provided me with the background and notions necessary for this Bachelor's thesis. My grateful thanks are also extended to Axel Dahlberg for taking the time and effort to help me with any questions related to SimulaQron. Finally, I would like to thank Michael Walter for taking the time to read my work.

# I. Introduction

By 2020, the Dutch research centre QuTech aims to have built a quantum network between four cities in the Netherlands: Amsterdam, Delft, Leiden, and The Hague [ALA18; DW18; QuT]. The ultimate goal is to construct a large-scale quantum network so that quantum information can be exchanged between remote quantum processors, allowing for both quantum computation and quantum communication over an arbitrarily long distance. Such a large-scale quantum-communication network is also known as a quantum internet [Cas18; DW18; DLH17]. Despite the fact that many possible applications of a quantum internet have been suggested in research (e.g. see [Weh08]), the real potential of a quantum internet will only be known when it has been realised. Nevertheless, the presumed arrival of a quantum internet will require software that can be run over the quantum network. Researchers have therefore created a quantum internet simulator – called SimulaQron – with the purpose to function as a test-bed for writing software for quantum internet applications [ALA18; DW18]. One of the potential applications of a future quantum internet is two-party cryptography [DW18; Weh08]. Cryptography is concerned with the tasks of achieving communication or computation between two or more parties, even if the parties do not trust each other [NC10]. A typical example of cryptography is that of two parties who want to communicate securely over a communication channel. We often refer to these two parties as 'Alice' and 'Bob', and to a (third) party attempting to intercept their messages as the 'adversary' or 'attacker'. A cryptographic protocol then describes how Alice and Bob can perform their task securely. In classical cryptography, the security of such cryptographic protocols is often based on the hardness of an unproven mathematical conjecture and on the assumption that the attacker has limited computational power [DLH17; KWW12]. However, the security of these protocols could be affected if the attacker obtains more (e.g. quantum) computational power. For instance, the algorithm developed by Shor [Sho95] for integer factorisation on a quantum computer could break RSA [RSA78], a cryptoscheme widely used for securing the transfer of (sensitive) data. It is therefore desirable to aim for protocols that are information-theoretically secure, that is, the security of the protocol can be proven without restrictions on the attacker's computational power [Sca+09].

Whilst the laws of quantum mechanics may allow for breaking currently used classical cryptoschemes, these laws could also be exploited to design quantum protocols for cryptographic tasks, which is the purpose of quantum cryptography. In fact, information-theoretical security can be achieved in the case of quantum-key distribution (QKD)[1], which enables two parties to establish a shared secret key. Nevertheless, not all cryptographic tasks can be performed with information-theoretical security, even with the help of quantum mechanics. This limitation holds, in particular, for secure two-party computation, which involves cryptographic tasks in which two parties want to exchange information but do not trust each other.[2] More precisely, it has been shown that additional assumptions on the attacker are inevitable in order to allow for secure two-party computation [BCS12; Lo97; LC97; May97]. These assumptions need to be proven as secure and realistic to effectively restrict the attacker's actions.

Consequently, much research has been devoted to designing cryptographic models for secure two-party computation protocols which introduce such additional assumptions on the adversary. Two cryptographic models that have been proposed are based on physical ( as opposed to computational) assumptions, mimicking the technical difficulties that one encounters when storing quantum information: the bounded-quantum-storage model [Dam+05; Dam+07] and the noisy-storage model [KWW12; Sch10; STW09; Weh08]. Whereas the first model gives an explicit upper bound on the adversary's quantum storage during the protocol, the latter assumes that the quantum storage is noisy [KWW12; Sch10; STW09; Weh08], meaning that stored (quantum) information goes lost over time. The noisy-storage model is perceived as more realistic compared to the bounded-quantum-storage model as it reflects more accurately the prevailing technical difficulties of storing quantum information [KWW12; Sch10], hence the paper will predominantly focus on the noisy-storage model.

Several protocols for secure two-party computation have been designed for the noisy-storage model, among which protocols for the primitive called 1-2 oblivious transfer (1-2 OT), such as provided in [KWW12] and [Sch10]. 1-2 OT is particularly interesting as it has been shown that any protocol for secure two-party computation can be based on this primitive [GV88; Kil88]. 1-2 OT is a cryptographic primitive in which Alice has two secret strings (messages, indicated by m) m0, m1 and Bob has a bit c $\in$ 0,1 and the goal is that Bob receives mc corresponding to his choice bit c, but Alice does not learn Bob's choice c; likewise, Bob does not learn Alice's remaining string [GV88; KWW12]. The protocols for 1-2 OT in the noisy-storage model require Alice and Bob to be connected by a quantum channel. Since it is realistic to assume that in any practical implementation of 1-2 OT such a quantum channel may be unreliable and hence induce errors on the transmitted quantum information, protocols have also been suggested that are robust against errors caused by the quantum channel, such as the one in [Sch10]. In particular, both the idealised and the robust protocol provided in [Sch10] are shown to be information-theoretically secure against any noisy-storage attack, as long as the laws of quantum mechanics hold and the amount of noise in the quantum storage channel is above a particular level.

Since secure two-party computation is one of the expected applications of a future quantum network [Weh08] and any secure two-party computation primitive can be based on 1-2 OT, it is significant to evaluate the implementation of a protocol for 1-2 OT over such a quantum network. In particular, since SimulaQron has been designed to de-

[1] For example, see [TL17] for a security analysis for QKD protocols.

[2] Note that, contrary to the previous example and QKD, Alice and Bob now do not trust each other, and thus the "adversary" can be either of the two. A concrete application of secure two-party computation is the so-called millionaire's problem in which Alice and Bob want to find out who is the richest of the two without revealing to the other how much money he or she has [Yao82].

velop software for future quantum-internet applications, it is important to verify its suitability for testing cryptographic tasks. Nonetheless, there are, to the best of our knowledge, no implementations of any cryptographic protocols using SimulaQron. Therefore, the present paper aims to implement both the idealised and the robust protocol for 1-2 OT from [Sch10] in SimulaQron in order to (i) examine whether the selected protocols are adequate for quantum internet applications, and (ii) evaluate the extent to which SimulaQron is a suitable environment for implementing and analysing cryptographic protocols.

## Methodology

The present research consists of a critical and extended literature review, followed by an implementation of the two protocols from [Sch10] using SimulaQron [Sim].

By critically analysing the existing literature on 1-2 OT in the noisy-storage model, this paper attempts to justify the choice for the protocols in [Sch10] and to be able to provide a deeper understanding of the protocol for implementation using SimulaQron. Moreover, in order to clarify the extra steps for error-correction for the implementation of the robust protocol in an unreliable quantum channel, it was necessary to study the error-correction techniques that are available. An analysis of the existing literature on quantum internet and SimulaQron was then used for understanding the working of SimulaQron and to evaluate the potential of SimulaQron.

SimulaQron has been chosen as the environment for our implementations as it allows for a simulation of the quantum network between Alice and Bob and hence enables (simulated) quantum communication to be performed, which is required for implementing the protocols. In order to carry out the implementations, we have used several tools provided by the Python Classical-Quantum Combiner (CQC) library that comes with SimulaQron and made the implementations feasible. We have written our code in Python [Pyt] and hence used several tools provided by Python and its libraries. For the decoding part, we use the source code from [Fil], which is also written in Python.

## Knowledge utilisation

This research provides new insights into the area of quantum internet development. Since a quantum internet has not yet been realised, the possible applications of a future quantum internet are still speculative. Therefore, by implementing and evaluating one of the presumed quantum-internet applications (i.e. cryptography) in a simulated quantum network, our research contributes to developing a deeper understanding of the suitability of a quantum internet for such cryptographic applications.

Moreover, since there are, to the best of our knowledge, no previous implementations of cryptographic applications using the recently developed SimulaQron, the results of this research explain the potential of SimulaQron for developing such applications. In addition, we shed light on the possibilities and current limitations of SimulaQron in general, which is of use for current and future users of SimulaQron and may contribute to a further development of this quantum-internet simulator. In fact, we may in this way also contribute to the development of the quantum network in the Netherlands that is planned to be realised in 2020, since a version of the CQC interface[3] used by SimulaQron is intended to be integrated into this quantum network as well.

Lastly, this research contributes to the field of quantum cryptography as there are, according to the literature, no quantum-network implementations of the protocols in [Sch10]. The only implementations of these and similar 1-2 OT protocols for the noisy-storage model are achieved by using hardware for QKD, such as done in [Erv+14]. Furthermore, since our code is written in the high-level language Python and publicly available on GitHub, our code can be used to provide an illustration of the protocols for 1-2 OT for learning objectives.

## Outline of the paper

The remaining part of the paper is organised in the following way. Section 2 describes some terminology of quantum mechanics used in this paper. Section 3 provides an analysis of the noisy-storage model and a description of the protocol for 1-2 OT in the idealised setting. Section 4 addresses how to correct errors in the setting of an unreliable quan-

[3] See Section 5.

tum channel and explains the resulting robust protocol for 1-2 OT. Section 5 presents a background on quantum internet and SimulaQron. The results of our implementations are provided in Section 6, which is followed by a discussion. Finally, we conclude our work and provide some potential directions for future research in Section 7.

## II. Background to quantum theory

In this section, we provide a brief introduction to quantum mechanics and focus on the concepts that are relevant for the remaining part of this thesis. We assume that the reader is familiar with the basic notions of linear algebra. We use the definitions provided in [NC10].

### Qubits

Recall that a classical binary bit either has value $0$ or $1$. The quantum version of the bit is often called a qubit, short for quantum bit. Similar to a classical bit, a *qubit* can be in state $|0\rangle$ or $|1\rangle$, where $|\rangle$ is the frequently-used *Dirac* notation. However, the main difference is that a qubit can be in any linear combination of these two states.

More precisely, we define a qubit as a vector in a two-dimensional complex vector space $\mathbb{C}^2$ with inner product, also called a Hilbert space. The states

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{and} \qquad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

form an orthonormal basis for the Hilbert space, often referred to as the *computational basis*. We write the most general quantum state as: $|\psi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle$, where $\alpha_0, \alpha_1 \in \mathbb{C}$ are normalised so that $|\alpha_0|^2 + |\alpha_1|^2 = 1$. Such a state is often called a *superposition* of states $|0\rangle$ and $|1\rangle$. Moreover, we call $\alpha_i$ the amplitude of state $|i\rangle$ for $i \in \{0, 1\}$.

### Measurement postulate

Qubits act in a probabilistic manner when measured. According to the Copenhagen interpretation of quantum mechanics, measuring a quantum state irrevocably changes its state: after measuring an arbitrary quantum state $|\psi\rangle$, if the outcome is $i$, then $|\psi\rangle$ is *collapsed* (i.e. changed) to state $|i\rangle$. Clearly, if the qubit is in state $|\psi\rangle = |0\rangle$ (i.e. $\alpha_0 = 1$, $\alpha_1 = 0$), then we get outcome $0$ with probability $1$. However, it may seem less trivial if a qubit is in superposition: if the qubit is in state $|\psi\rangle = \alpha_0 |0\rangle + \alpha_1 |1\rangle$ with $\alpha_0$ and $\alpha_1$ both nonzero, we either get outcome $0$ with probability $|\alpha_0|^2$ or $1$ with probability $|\alpha_1|^2$.

### BB84 states

The following four states are often referred to as the BB84 states, since these states were used in the BB84 protocol **?** for QKD (named after the authors Charles Bennett and Gilles Brassard and the year in which it was proposed).

$$|0\rangle \qquad |1\rangle \qquad |+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \qquad |-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$$

As aforementioned, $\{|0\rangle, |1\rangle\}$ is called the computational basis. We will refer to $\{|+\rangle, |-\rangle\}$ as the *Hadamard basis*.

### Operators

An operator acts on a quantum state and enables us to change the state of the corresponding qubit. An operator is defined as a linear transformation on the Hilbert space $\mathbb{C}^2$. By the laws of quantum mechanics, an operator $U$ must be unitary, i.e. $U^\dagger U = I$, where $I$ represents the identity, and $U^\dagger$ means the conjugate transpose of $U$. Below, we provide two examples of operators, which will also be used in the implementations of this paper.

The first operator we consider is denoted $X$. It acts as follows on the BB84 states:
$X|0\rangle = |1\rangle$, $X|1\rangle = |0\rangle$, $X|+\rangle = |+\rangle$, and $X|-\rangle = -|-\rangle$. In other words, if the qubit is in one of the computational basis states, the $X$ operator swaps the states, whereas if the qubit is in one of the Hadamard basis states it remains in the same state (with $ket-$ mapped to $-|-\rangle$). Note the connection to the logical NOT gate.

An operator can be represented by a matrix. The matrices corresponding to $X$ and $H$ are:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

## III. 1-2 oblivious transfer in the noisy-storage model

As mentioned before, 1-2 oblivious transfer (1-2 OT) is a cryptographic task belonging to the field of secure two-party computation on which any other protocol for secure two-party computation can be based [GV88; Kil88]. Oblivious transfer was first introduced by Rabin [Rab81] and has been generalised to 1-2 OT by Even, Goldreich, and Lempel [EGL82]. We recall that in the setting of 1-2 OT, Alice has two secret strings $m_0$, $m_1$ and Bob chooses a bit $c \in \{0, 1\}$. The task that 1-2 OT needs to accomplish is that Bob receives Alice's input string $m_c$ corresponding to his choice bit $c$, without revealing $c$ to Alice and without learning about Alice's other string. Note that here, and in the remaining part of this thesis, 'Alice' and 'Bob' refer to the two parties that want to achieve the task of 1-2 OT, and that the 'attacker' can be any of these two parties (recall footnote 2).

## Additional assumptions

We explained in Section 1 that even with the possibility of quantum communication, protocols for secure two-party computation—and hence 1-2 OT—must impose additional assumptions on the adversary in order to achieve a desired level of security [BCS12; Lo97; LC97; May97]. It is reasonable to aim for protocols for secure two-party computation that are feasible and based on realistic assumptions. Firstly, feasibility is important to ensure that the protocol can be used in practice. Moreover, the assumptions should be realistic in the sense that they take into account the state-of-the-art technologies and their limitations [KWW12; Sch10]. Therefore, two cryptographic models have been proposed that are based on physical (instead of computational) assumptions: the bounded-quantum-storage model [Dam+05; Dam+07] and the noisy-storage model [WST08; STW09; Sch10; KWW12].

### Bounded-quantum-storage model

The physical assumption in the bounded-quantum-storage model is that the adversary's quantum memory size is limited to a certain number of qubits. No restriction is imposed on the adversary's classical memory or computing power. In this model, the version of OT that was introduced in [Rab81] was proven to be information-theoretically secure [Dam+05]. In particular, if Alice and Bob are honest, they do not need any quantum memory and the protocols can only be broken if the dishonest party has a quantum memory of size at least $\frac{n}{2}$, where $n$ is the number of qubits transmitted during the protocol [Dam+05]. Moreover, the findings of [Dam+05] have been extended to the case of 1-2 OT: 1-2 OT is shown to be secure against a dishonest player whose quantum memory is of size at most $\frac{n}{4} - 2l$ [Dam+07]. Here, $n$ amounts again to the number of qubits that are transmitted during the protocol, and l (which may be a fraction of n) is the length of Alice's input strings $m_0$, $m_1$. In addition, [Dam+05; Dam+07] point out that their protocols also provide security in the case of imperfect quantum communication, a situation that will be covered in Section 4.

There are three reasons why we may consider the bounded-quantum-storage model. Firstly, if Alice and Bob merely communicate over a classical channel—and thus not use quantum communication—secure 1-2 OT can only be achieved if the adversary's classical memory size is at most quadratic in memory size of an honest party [DM04]. However, using quantum communication, the protocols for the bounded-quantum-storage model only require a bound on the size of the adversary's quantum memory. In particular, these protocols achieve a desired level of security regardless of the size of the adversary's classical memory, which, consequently, allows for a bigger ratio between the memory size of the honest parties and that of an adversary than would be possible in the classical setting. In fact, the honest parties do not need any quantum storage at all, and the bound on the adversary's *quantum* memory size only depends on the number $n$ of qubits that have been sent during the protocol. Therefore, the bounded-quantum-storage model allows us to outdo the optimal bound for which security is obtained in the case that a dishonest player only has access to a classical memory [Dam+05]. Secondly, the limited size of the adversary's quantum memory ensures that a dishonest party will lose information: being unable to store all qubits, the adversary must store some information in his classical memory (i.e. by measuring the qubits) [Dam+05]. It is, in fact, this irreversible loss of information that allows for 1-2 OT in the bounded-quantum-storage model [Dam+05]. The third reason to consider the bounded-quantum-storage model is a practical one: whereas quantum transmission and quantum measurements — needed for the honest parties — are possible with the present state-of-the-art technology, storing only “a single qubit for more than a fraction of a second” is technologically still very difficult [Dam+05], impeding the storing possibilities of a dishonest party.

Despite the aforementioned reasons, it is questionable whether any practical situation will impose such a limit on the quantum memory size as the bounded-quantum-storage model requires [WST08]. In fact, [Dam+05] argue that it is more realistic to consider the case that the adversary’s quantum memory is noisy, instead of bounded, where noise induces loss of (stored) quantum information. Therefore, we describe the alternative model and its advantages below.

### Noisy-storage model

Whereas the bounded-quantum-storage model gives an explicit upper bound on the adversary’s quantum memory size during the protocol, the noisy-storage model[4] assumes that the quantum storage (i.e. the quantum memory) is noisy [KWW12; Sch10; STW09; WST08]. More precisely, this model allows an adversary to have the best available quantum memory, but this memory is affected by noise over time: the longer the adversary stores quantum information, the higher the amount of prevailing noise, and hence the more information is lost [Sch10]. Two-party protocols for the noisy-storage model (such as the protocols considered in this paper) make use of this time-dependency by inducing a waiting time for the participating parties during the protocol. After this waiting time, the higher level of noise in the adversary’s quantum memory results in information loss, see Section 3.2. As a consequence, we can evaluate the security of a protocol for this model in terms of the noise level [WST08]. Furthermore, note that the bounded-quantum-storage model can, in fact, be obtained from the noisy-storage model by simply assuming that there is no noise in the quantum memory but that the input of the quantum memory (i.e. the number of qubits that can be stored) is limited [KWW12].

By exploiting the technological difficulty of building a quantum computer (i.e. the presence of noise) the noisy-storage model has significant practical value. In fact, the noisy-storage model reflects the prevailing technical difficulties of storing quantum information more accurately than the bounded-quantum-storage model. First of all, current quantum memories are not completely reliable. In particular, with current technology, the time during which quantum information can be stored without being completely lost is limited (e.g. see [Zho+12]). The imperfection of current quantum memories results from the difficulty to keep the stored quantum information unaltered over time and the problem that quantum information (i.e. the photons in which the quantum information is encoded) is lost during storage [Erv+14]. In fact, [KWW12] argues that it is not even known whether it is physically possible to build a fully reliable memory in the future. In addition, noise may arise when quantum information is transferred onto a different physical carrier representing the quantum memory; for example, when photonic qubits are transferred onto an atomic ensemble or an atomic state [STW09]. In fact, this transfer onto the quantum memory is often already noisy [KWW12], irrespective of the (im)perfection of the quantum memory itself [Weh+10].

Yet, we should note that security is only guaranteed if the aforementioned technological challenges prevail [Erv+14]. Nevertheless, the noisy-storage model is resistant against technological improvements in the future (i.e. security can still be obtained) as long as the quantum memory size has a finite upper bound [Erv+14]. This resistance of the model results from the fact that the level of security depends on the number of qubits that are transmitted during the protocol, and hence can be monitored. In other words, if the dishonest party has access to a larger quantum storage, the required level of security can still be obtained by sending more (quantum) information during the protocol [Erv+14]. Hence, as long as an adversary does not have access to an unlimited and noiseless quantum memory in the future—which can be perceived a reasonable assumption – we conclude that the noisy-storage model is a realistic and feasible model for cryptographic purposes.

[4] Since the noise occurs in the quantum memory, researchers sometimes also call this model the noisy-quantum-storage model. However, in this thesis we stick to the more common noisy-storage model. In particular, we assume in this thesis that there is no noise in the classical communication channel and classical storage at all.

## Security in the noisy-storage model

Given the aforementioned reasons why the noisy-storage model could be seen as a realistic reflection of the prevailing technical difficulties, we based our implementations on protocols for 1-2 OT within the noisy-storage model[5]. As previously explained, the noisy-storage model allows a dishonest party to have access to an unlimited quantum memory, yet this quantum memory is noisy. This section provides an intuitive description of what such a noisy quantum memory entails for a dishonest party and why this may allow us to achieve security. We also describe another tool, *privacy amplification*, that is used in the considered protocols to provide security in the noisy-storage model.

### Noisy-quantum-storage attacks

First of all, note that in the considered protocols for 1-2 OT in the noisy-storage model only a dishonest Bob (i.e. the receiver) can take advantage of his quantum memory, since he is the only one that receives quantum information. In particular, security against a dishonest Alice does not depend on her quantum memory because the only information she receives from Bob is classical information. Therefore, we only consider noisy-quantum-storage attacks by a dishonest Bob. Suppose now that Bob tries to store an incoming quantum state in his quantum memory. Since the quantum memory is noisy, the state is affected by noise and decoheres over time [KWW12], i.e. the quantum system interacts with its environment and information is lost. Then, security in this model can be achieved by imposing a waiting time during the protocol: the quantum information that a dishonest Bob may store in his quantum memory undergoes noise during this waiting time, which results in loss of information. In this way, we may exploit the noise in the quantum memory in order to achieve the desired level of security. Nevertheless, we emphasise that a dishonest Bob is still allowed to perform any encoding attack on the received information (i.e. the attacker has unlimited computational power and can do perfect quantum operations) and has an unlimited noiseless classical storage. Also, after the waiting time, Bob can do any decoding attack. See Figure 1 for an illustration of a noisy-quantum-storage attack by dishonest Bob.

### Privacy amplification

We provide an intuition about the tool of privacy amplification [BBR88], which is applied during the protocol (i.e. Step 5 of Protocol 3.1) to ensure that any leaked information to a dishonest party is removed, and thus security can be guaranteed. Suppose that Alice holds a random bitstring $A$ of length $n$ about which a dishonest Bob has partial information.This partial information may allow Bob to guess $A$, since $A$ does not look completely random to Bob anymore. Then privacy amplification is performed using some randomly chosen two-universal hash function. In short, a hash function is a function that maps a set of some size to a smaller set [CW79]. We say that a class $\mathcal{F}$ of functions $f : \{0,1\}^n \to \{0,1\}^l$. is two-universal if $P_{f \in_R \mathcal{F}}[f(x) = f(x')] \leq s^{-l}$ for all $x \neq x' \in \{0,1\}^n$ [CW79]. Note that $f \in_R \mathcal{F}$ means that $f$ is randomly chosen from $\mathcal{F}$. Then, by applying a randomly chosen two-universal hash-function to her $n$-bit string $A$, Alice obtains a shorter string, about which Bob has a negligible amount of information. In particular, the probability that Bob guesses Alice's shorter string is almost uniformly random [Erv+14]. In other words, by shrinking string $A$ in this way, privacy amplification allows for removing the leaked information and thus increases the level of security.

## Protocol for 1-2 OT

Several protocols for 1-2 OT in the noisy-storage model have been suggested and research distinguishes two types of attacks from the adversary in their security proofs: individual and general storage attacks. When the adversary is allowed to only perform measurements on each incoming qubit individually, we call this an individual attack. However, since multi-qubit measurements are possible with current technology, it is more realistic to assume that the attacker can perform any arbitrary attack, i.e. a general attack [KWW12]. Whereas the protocols in [WST08] and [STW09] are proven to be secure only against individual storage attacks, the protocols in [KWW12] and [Sch10] are proven to be

[5] For a formal definition of the noisy-storage model, we refer the reader to [Weh+10].

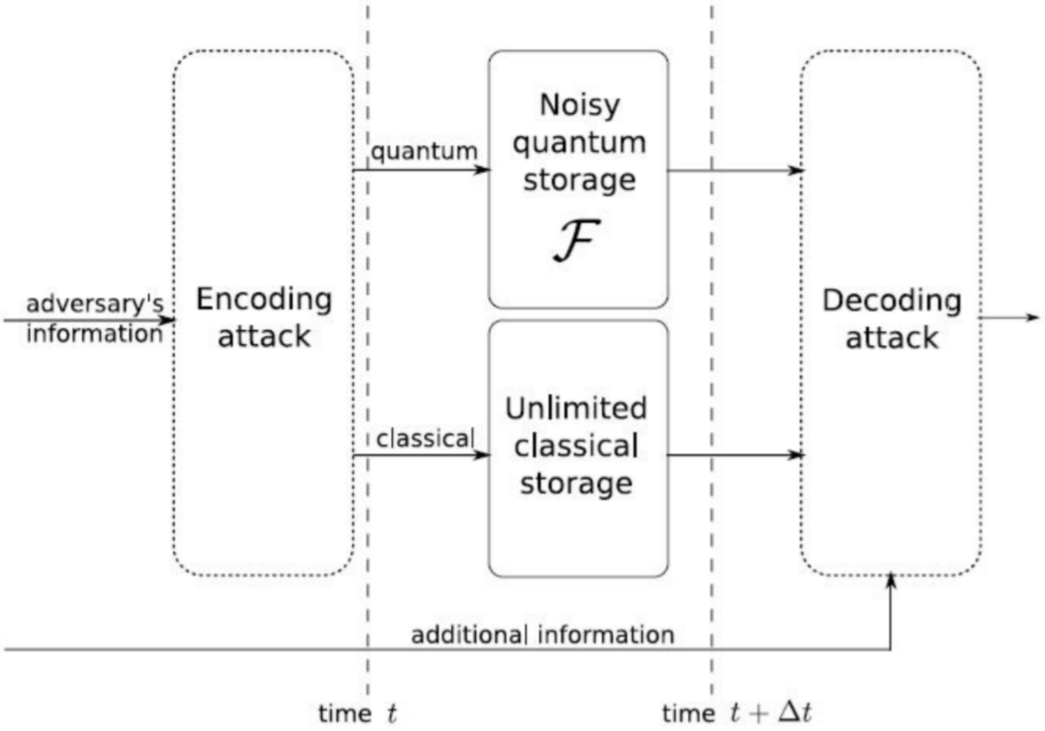


Figure 1: (Source: [Weh+10]) An illustration of an attack by dishonest Bob. Here, $\mathcal{F}$ represents the noisy-quantum memory (as formally defined in [Weh+10]) and $\Delta t$ the waiting time during the protocol.

secure against any attack. The latter two protocols are hence much more realistic to consider.

In particular, the protocols in [KWW12] and [Sch10] have been proven to be information-theoretically secure against general noisy-storage attacks as long as the laws of quantum mechanics hold and the amount of noise in the quantum storage channel is above a particular level. Recall from Section 1 that an information-theoretically secure protocol implies that there are no restrictions on the attacker's computational power, which is in contrast to many classical cryptographic protocols that are based on computational assumptions and hence are more restricted in their performance. Although the protocol in [KWW12] requires less storage noise for security, the protocol in [Sch10] is more straightforward and, therefore, simpler to implement. For this reason, we have decided to implement the protocol given in [Sch10], which will be covered in Section 6.

**Protocol for 1-2 OT in the noisy-storage model**

Before we provide Protocol 3.2 for 1-2 OT in the noisy-storage model, we provide a protocol for *randomised* 1-2 OT (1-2 ROT), because, as will be further clarified, 1-2 OT can be easily obtained from 1-2 ROT [Sch10]. In 1-2 ROT, Alice has no input whilst Bob holds his choice bit $c \in \{0, 1\}$. After completion of the protocol, Alice holds two bit-strings of the same length, $s_0$ and $s_1$, and Bob holds $s_c$, i.e. the string received by Alice corresponding to his choice bit $c$. See Figure 2 for an illustration.

We first explain some notation, for which we use some terminology from Section 2. Note that before running Protocol 3.1 and Protocol 3.2, Alice and Bob agree on the length $\ell$ of the strings $s_0$ and $_1$. The level of security[6] is also decided beforehand, and determines the number $n$ of qubits that will be transmitted in Step 1. Moreover, note that $X \in_R Y$ means that $X$ is *randomly* picked from $Y$. We also note that $\mathrm{H}$ in Step 1 is the Hadamard operator, and thus if $y_i^A = 1$, Alice first applies $\mathrm{H}$ to bit $x_i^A$ (for $i \in [n]$) before she sends it to Bob. For clarification, this implies that in Step 1, Alice randomly sends one of the four BB84 states ($|0\rangle, |1\rangle, |+\rangle, |-\rangle$) to Bob. Finally, in Step 2, when Bob measures his incoming qubit in the Hadamard basis, we mean that Bob first applies $\mathrm{H}$, and then measures in the computational basis.

[6] With the level of security we refer to the security error-probability $\epsilon$ for an $\epsilon$-secure 1-2 ROT protocol as defined in [Sch10].

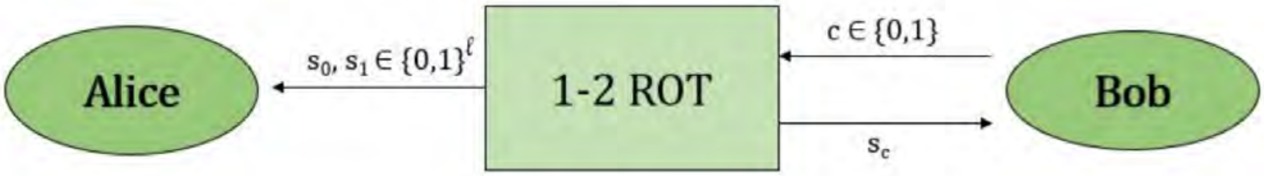


Figure 2: In the setting of 1-2 ROT, Bob inputs a bit $c \in \{0, 1\}$. After completion, Alice holds $s_0, s_1 \in \{0, 1\}^l$ and Bob holds $s_c$, which is one of Alice's output strings. Before running the protocol, Alice and Bob agree on the length $\ell$ of the output strings.

The Protocol below (Protocol 3.1) is the protocol for 1-2 ROT as provided in [Sch10]. This protocol originates from [Ben+92b; Dam+09], and security of this protocol in the noisy-storage model is proven in [Sch10].

[Ben+92b; Dam+09] 1-2 ROT$^l$

1. Alice randomly picks $x^A \in_R \{0,1\}^n$ and $y^A \in_R \{0,1\}^n$.
   At time $t = 0$, Alice sends $\mathrm{H}^{y_1^A}|x_1^A\rangle, ..., \mathrm{H}^{y_n^A}|x_n^A\rangle$ to Bob.

2. Bob randomly picks $y^B \in_R \{0,1\}^n$.
   Bob measures the $i$-th qubit he receives in the computational basis if $y_i^B = 0$ and the Hadamard basis if $y_i^B = 1$. He obtains outcome $x^B \in \{0,1\}^n$.

Both parties wait time $\Delta t$.

3. Alice sends her basis string $y^A$ to Bob.

4. Bob forms the sets $I_c = \{i \in [n] \mid y_i^A = y_i^B\}$ and $I_{\bar{c}} = \{i \in [n] \mid y_i^A \neq y_i^B\}$, where $c$ is his choice bit and $\bar{c}$ the remaining bit.
   Bob sends $I_0$ and $I_1$ to Alice.

5. Alice picks two hash functions $f_0, f_1 \in_R \mathcal{F}$, where $\mathcal{F}$ is a class of two-universal hash functions $f : \{0,1\}^n \to \{0,1\}^l$.
   Alice sends $f_0$ and $f_1$ to Bob and outputs $s_0 = f_0(x^A\restriction_{I_0})$ and $s_0 = f_1(x^A\restriction_{I_1})$.[7]

6. Bob outputs $s_c = f_c(x^B\restriction_{I_c})$.

Note that, after Step $2$, whenever $y_i^A = y_i^B$ for $i \in [n]$, we also have that $x_i^A = x_i^B$, since Bob measures his $i$-th incoming qubit in the same basis as in which Alice encoded her $i$-th bit.

As pointed out in [Sch10], 1-2 OT can be achieved from 1-2 ROT. Recall that in the setting of 1-2 OT, Alice has two input bit-strings of length $\ell$, say $m_0$ and $m_1$, and Bob|holding his choice bit $c \in \{0,1\}$|wants to obtain $m_c$ after completion of the protocol. Below we provide the resulting protocol for 1-2 OT. An illustration of the required steps is provided in Figure 3. We emphasise that a quantum channel between Alice and Bob is only required for achieving 1-2 ROT, i.e. in Step $1$. Moreover, we use the same notation as for the preceding protocol and $\ell$ and $n$ are again determined beforehand.

### Protocol 3.2: 1-2 OT$^l$

1. Alice and Bob run the protocol for 1-2 ROT.
   After completion, Alice holds two uniformly random bit-strings $s_0$ and $s_1$ of length $\ell$, and Bob holds $s_c$ corresponding to his choice bit $c$.

2. Alice computes $c_0 = m_0 \oplus s_0$ and $c_1 = m_1 \oplus s_1$, and sends them to Bob over the classical channel.

3. Bob receives $c_0$ and $c_1$. Bob now obtains $m_c$ by computing $s_c \oplus c_c = s_c \oplus (m_c \oplus s_c) = m_c$.

For an explicit analysis of the correctness and a full proof of security of these protocol in the noisy-storage model, we refer the reader to [Sch10].

[7] If the length of $x^A\restriction_{I_0}$ or $x^A\restriction_{I_1}$ is $m < n$, then Alice adds $n - m$ $0$'s to the string before applying $f_0$ or $f_1$, respectively.

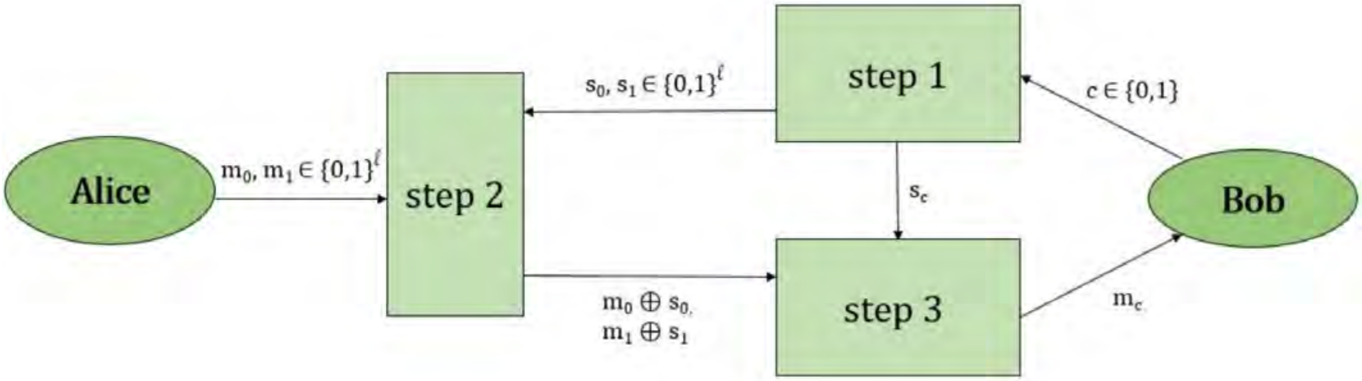


Figure 3: In the setting of 1-2 OT, Alice inputs two strings $m_0, m_1 \in \{0,1\}^l$ and Bob inputs a bit $c \in \{0,1\}$. The three indicated steps are those needed to achieve 1-2 OT from 1-2 ROT. After completion, Bob holds $m_c$, Alice's input string corresponding to his choice bit $c$. Contrary to the setting of 1-2 ROT, Alice now does not have an output.

## IV. Error correction and robust 1-2 OT

In the practical implementation of 1-2 OT it is realistic to assume that the quantum channel connecting Alice and Bob is subject to a particular level of noise. Specifically, imperfections occur when a quantum channel is implemented in practice, resulting in errors during the transmission of quantum information over the channel [Ben+92a; BS94; EMMM11; GPCZ18]. Apart from the transmission of the qubits itself, the preparation of the qubits by Alice and the measurements by Bob will also be affected by noise [KWW12]. Therefore, we should make the protocol for 1-2 OT *robust* against noise for the honest parties [STW09] so that the protocol is suitable for real implementations. In other words, we need to find a way to deal with the errors resulting from the noisy communication channel and the imperfections during the preparation and measurement of quantum states. This section, in particular, will describe how *information reconciliation* – an additional step during which error-correction techniques are applied – can provide a solution, and subsequently explain the resulting protocol for robust 1-2 OT in the noisy-storage model[8].

### Information theory

Before continuing, we formalise some notions of information theory that will be used in this section.

We use the definitions provided in [DS17]. We assume that the reader is familiar with the basic notions of probability theory; however, the relevant definitions are provided in Appendix A.

An important measure of uncertainty about information is the Shannon entropy. We provide three definitions below.

**Definition 1. (Shannon entropy)**
Let $X$ be a random variable on the set $\mathcal{X}$. We define the *(Shannon) entropy* of $X$ as

$$H(X) := -\sum_{x\in\mathcal{X}} P_X(x) \cdot \log_2 P_X(x).$$

In other words, $H(X)$ measures the amount of uncertainty about $X$: the higher the uncertainty about $X$, the higher $H(X)$. Note, however, that it is a function of $P_X$, not of $X$ itself.

**Definition 2. (Conditional entropy)**
Let $X$ and $Y$ be two random variables on the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, the *conditional entropy* of $Y$ given $X$ is the average uncertainty about $Y$

[8] Note that a noisy communication channel should not be confused with a noisy quantum memory (as in the noisy-storage model), as these notions have a completely different meaning. In particular, if Alice and Bob are both honest, they do not even need a (noisy) quantum memory in order to perform 1-2 OT; however, this does not imply that the quantum channel connecting Alice and Bob should be noiseless: errors can still occur and affect the qubits that are sent from Alice to Bob during the protocol.

when the outcome of $X$ is known, i.e.

$$H(Y|X) := -\sum_{y\in\mathcal{Y},x\in\mathcal{X}} P_{YX}(y,x)\cdot\log_2 P_{Y|X}(y|x).$$

**Definition 3. (Mutual information)**
The *mutual information* of two random variables $X$ and $Y$ is defined as

$$I(Y;X) := H(Y) - H(Y|X) = H(X) - H(X|Y).$$

In words, $I(Y;X)$ measures the reduction in uncertainty about $Y$ when $X$ is known, and vice versa.

To provide some intuition, suppose that $X$ is the input of a communication channel and $Y$ the output. Then the mutual information measures how much information output $Y$ reveals about input $X$.

**Definition 4. (Discrete channel)**
A *(discrete) channel* is a triple $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ consisting of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a conditional probability distribution $P_{Y|X} : \mathcal{Y} \to [0,1]$. Here, $\mathcal{X}$ represents the set of possible channel inputs and $\mathcal{Y}$ the set of possible channel outputs. $P_{Y|X}(y|x)$ is the probability of receiving output $y \in \mathcal{Y}$ when input $x \in \mathcal{X}$ is given.

**Definition 5. (Memoryless channel)**
A channel is called *memoryless* if the probability distribution of the output depends only on the current input.
In other words, if the memoryless channel is used multiple times, previous inputs and outputs do not affect the current output.

**Definition 6. (Noiseless channel)**
A channel $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ is *noiseless* if both $H(X|Y) = 0$ and $H(Y|X) = 0$.
In other words, the output of the channel completely determines the input, and vice versa.

## Noisy communication channel

Before we introduce the concept of information reconciliation and describe the protocol for robust 1-2 OT, we discuss in more detail the concept of a noisy communication channel. We consider the case that Alice sends a classical bit-string, $A$, to Bob over a communication channel. It is often assumed that the communication channel is reliable or noiseless, i.e. $A = B$, where $B$ is the bit-string received by Bob. However, in real implementations, Alice's string may be affected by noise (i.e. errors) during transmission so that $A \neq B$. In the presence of such noise, we say that the communication channel is *unreliable* or *noisy*. Clearly, if Alice and Bob transmit information over an unreliable communication channel, the correctness and security of the corresponding cryptographic protocol may be affected [GPCZ18].

A noisy communication channel can be modelled by a *binary symmetric channel* (BSC) [BS94]. A BSC is a communication channel in which each transmitted bit is flipped (i.e. affected by noise) with a certain probability $p$ [Mac03]. In other words, in the case that Alice sends her bit-string to Bob, each bit is transmitted correctly with probability $1-p$. Formally, i.e. according to Definition 4.4, we define a BSC with parameter $p$ by $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and by

$$P_{Y|X}(0|0) = P_{Y|X}(1|1) = 1-p,$$

$$P_{Y|X}(0|1) = P_{Y|X}(1|0) = p,$$

where $p \in [0,1/2]$ [DS17]. We will use BSC($p$) to refer to a BSC with probability parameter $p$. Note that a BSC is memoryless, as the probability that a transmitted bit flips (i.e. p) does not depend on any previous bit that has been transmitted over the channel. Moreover, note that the channel is noiseless if $p = 0$. An illustration of the BSC is provided in Figure 4.

## Information reconciliation

As mentioned before, the presence of noise requires an additional step during the protocol in which Bob can recover the information that Alice has sent so that the correctness of the cryptographic protocol can still be achieved. In the setting of 1-2 OT, this implies that Bob eventually receives Alice's input string $m_c$ corresponding to his choice bit $c$, even if the quantum channel is noisy.

During this additional step, the errors between the transmitted and received information are identified and error correction is performed. In the context of QKD, this process of finding the differences between the strings and performing error correction is often called *information reconciliation* or, simply, *reconciliation* (e.g. in [Ben+92a; BS94;

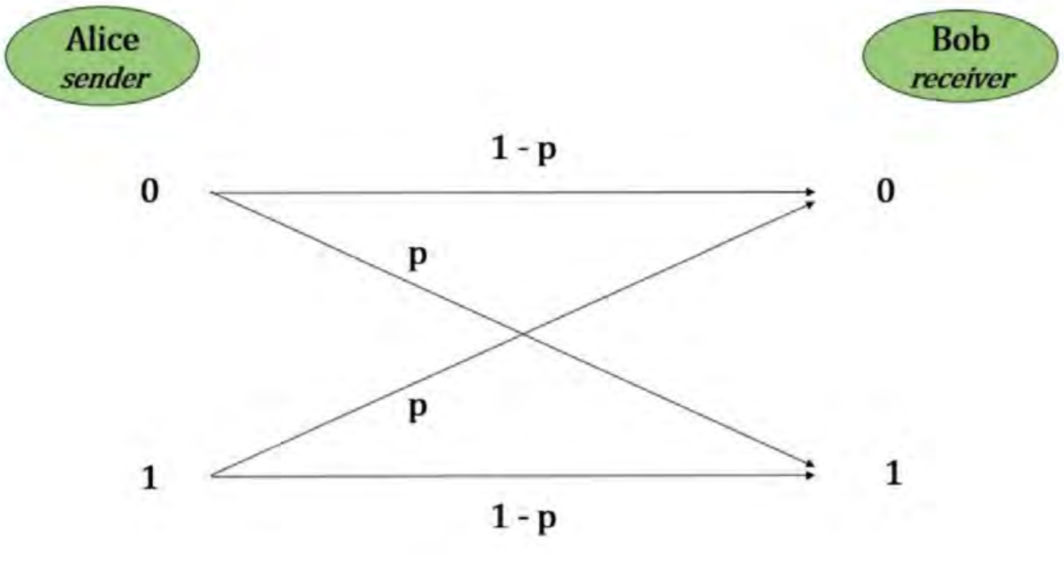


Figure 4: A binary symmetric channel between Alice and Bob, where each bit that Alice sends to Bob is flipped with probability $p$.

EMMM11]). Information reconciliation occurs via a public classical channel connecting the two parties [Ben+92a; EMMM11].

We provide a simple example of an information-reconciliation protocol by adjusting the description in [BS94] to the setting of 1-2 OT. Suppose again that Alice sends a bit-string $A$ to Bob over a noisy communication channel, and Bob receives the bit-string $B$. As $B$ is affected by noise, we may assume that $B \equiv A + N \pmod 2$, where the bit-string $N$ represents the noise. In the setting of 1-2 OT, an information-reconciliation protocol aims to generate an estimate, $\hat{A}$, of Alice's string $A$, given $B$. As we will see, this string $\hat{A}$ is obtained by exchanging some classical information over a public channel connecting Alice and Bob. At the end of the reconciliation protocol, Bob holds the string $\hat{A}$. Clearly, the protocol should minimise the amount of information revealed to a potential adversary and the probability that the protocol fails to produce the final string (i.e. $\hat{A}$) should be low [BS94].

### Privacy amplification

Nevertheless, it is important to note that by transmitting some extra information over a public channel, information reconciliation allows a dishonest party to gain additional information. Recall the tool of privacy amplification discussed in Section 3.2, which allows us to reduce the amount of information leaked to a potential dishonest party. In particular, since there is additional information leaked during the information-reconciliation protocol, we should perform more privacy amplification to achieve a desired level of security than would be the case if the quantum channel is noiseless and we need not perform information reconciliation. For example, suppose that there are m bits of additional information sent during information reconciliation, then the hash function should map the initial string to a string that is m bits shorter than in the case that the channel is noiseless.

### Error-correcting codes

Shannon [Sha48] described that reliable communication over a noisy channel can be obtained by means of encoding and decoding. We refer to an encoding and decoding system as an *error-correcting code*. Intuitively, we may describe encoding and decoding as follows. Instead of immediately sending her message $A$ to Bob, Alice first encodes her message into $T$ (the so-called *codeword*), by adding some *redundancy* to $A$ [Mac03]. Subsequently, $T$ is transmitted over the noisy channel, which outputs the received codeword $R \equiv T + N \pmod{}2$. Here, $N$ is a bit-string representing the noise induced by the channel. The decoder then uses the (known) redundancy in order to translate $R$ into the bit-string $\hat{A}$, Bob's estimate of $A$. Figure

5 provides an illustration.

### Shannon's noisy-channel coding theorem

t is important to mention that there is an upper bound on the amount of reliable communication that can be achieved in the presence of a noisy communication channel. This limit is known as Shannon's noisy-channel coding theorem [Sha48], and hence also referred to as the *Shannon limit*. Before we state the theorem, we provide two more definitions, obtained from [DS17].

**Definition 7. (Information rate)**
The *information rate* $R$ is the number of bits of information that are transmitted per channel use.

**Definition 8. (Channel capacity)**
Let $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ be a discrete, memoryless channel. Then its *channel capacity* $C$ is defined as

$$C := \max_{P_X} I(X;Y).$$

In other words, the channel capacity measures the maximum amount of mutual information between the input $X$ and the output $Y$ of a the channel.

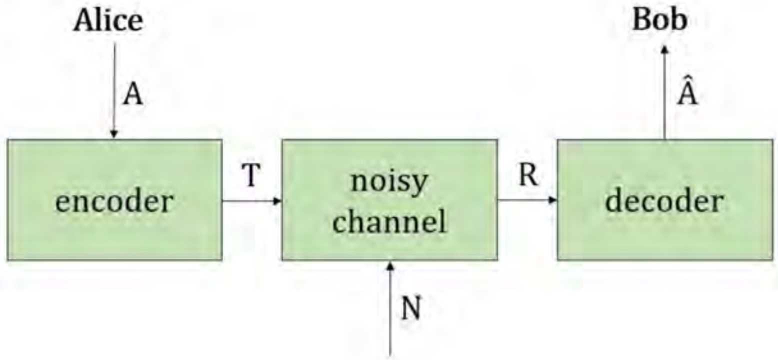


Figure 5: An illustration of an error-correcting communication system for the setting that Alice wants to send a message $A$ to Bob over a noisy communication channel.

The theorem below shows that information can only be transmitted with arbitrarily small error-probability over a given channel as long as the channel capacity is not exceeded. This is a slightly adapted version of the one provided in [DS17], and we refer the reader to this work for the proof.

**Theorem 1.** ***Noisy-channel coding theorem***
*Consider a discrete memoryless channel with capacity $C$. Then, (i) for any rate $R < C$, there is a method of encoding such that information can be transmitted over the channel with negligible maximal error-probability (i.e. the maximal error-probability approaches $0$ as the number of channel uses approaches infinity); and (ii) for any rate $R > C$, it is impossible to find a coding system that achieves such a negligible average error-probability.*

In other words, whilst the channel capacity $C$ indicates the maximal information rate that *theoretically* can be realised, Theorem 1 shows that it should also be possible *in practice* to achieve any rate $R < C$. We therefore aim to find a coding system that approaches the Shannon limit as much as possible.

## Robust protocol for 1-2 OT

In order to make our protocol for 1-2 OT robust against noise for the honest parties, we should add the extra step of information reconciliation. For our implementation of reconciliation in the robust 1-2 OT protocol, we use a class of error-correcting codes called Reed-Solomon (RS) codes. Before we explicitly describe how information reconciliation works in the setting of 1-2 OT and provide the robust protocol, we briefly introduce the class of RS codes.

### Reed-Solomon codes

We provide a high-level description of RS codes obtained from [LC83] and refer the reader to this work for a more explicit explanation. RS codes are a class of codes that were introduced by Reed and Solomon in 1960 [RS60]. Each message consists of elements from the Galois (i.e. finite) field GF($p^r$), for $p$ prime. If the original message has length $n$ and the transmitted codeword length $m$, then there are $m-n$ redundant symbols, and an RS code can correct up to $\frac{m-n}{2}$ errors.

In the following we briefly explain how encoding and decoding of RS codes is established, and use, for clarity, the same notation as in Figure 5. Each message $A$ of length $n$ is associated with a polynomial of the same length. Then, the transmitted codeword obtained after encoding can be considered as a polynomial $T(x)$ of some length $m > n$. The noisy communication channel then adds some noise $N(x)$ (also called the *error pattern* in [LC83]) to $T(x)$. Hence, we can decompose the received message as $R(x) = T(x) + N(x)$. If all coefficients of $N(x)$ are zero, then there are clearly no errors. However, if there are errors, the information from the received message $R(x)$ is recovered by the decoder. In short, the decoding process consists of four steps [LC83]. First, the *syndrome* of $R(x)$ is computed, which, intuitively, is an indicator of the error pattern (in particular, if the syndrome is zero, we assume that there are no errors). Then, from this syndrome we may determine the so-called *error-location polynomial*. In the third step, the error locations are retrieved by computing the roots of the error-location polynomial, and thus the error pattern (i.e. $N(x)$) is obtained. In the last step, we compute $R(x) - N(x)$ to obtain the transmitted codeword $T(x)$.

### Information reconciliation for 1-2 ROT

Having provided some intuition into RS codes, we can now give a complete description of the setting of randomised 1-2 OT when the quantum channel is noisy. Although the channel is noisy, we assume that the noise in the quantum channel is independent and identically distributed (i.i.d.), so each qubit that is transmitted over the channel is affected by noise with the same probability.

Recall from Protocol 3.1 for 1-2 ROT that Alice randomly picks two $n$-bit strings $x^A$ and $y^A$, and Bob randomly picks an $n$-bit string $y^B$. In this section, we consider the case that Alice sends each bit $x_i^A$ encoded as $\mathrm{H}^{y_i^A}|x_i^A\rangle$ to Bob over a *noisy* quantum channel. Bob measures each received qubit either in the computational ($y_i^B = 0$) or Hadamard ($y_i^B = 1$) basis . We now explain what may go wrong in the noisy setting. Recall that if the channel was noiseless, then for each $i$ such that $y_i^A = y_i^B$ (i.e. Alice sends the qubit in the same basis as in which Bob will measure) we have that $x_i^A = x_i^B$ with probability 1. However, if the channel is noisy, then the noise in the channel may affect the qubits that Bob has received. Therefore, when Bob measures each received qubit in basis corresponding to $y_i^B$, it is possible that his outcome $x_i^B$ is no longer such that $x_i^A = x_i^B$.

Now, recall that at a certain point during the protocol Bob forms the set $I_c$ containing all indices $i$ such that $y_i^A = y_i^B$, where $c$ is his choice bit. [9] Thus, in the presence of noise, it is possible that $x^A\restriction_{I_c} \neq x^B\restriction_{I_c}$, whereas equality would necessarily hold if the qubits were transmitted over a noiseless channel. Consequently, classical error-correction techniques need to be applied so that Bob obtains an estimate for $x^A\restriction_{I_c}$.

In order to understand how information reconciliation works in this setting, we explain how we can represent the occurring noise. For simplicity, let $A = x^A\restriction_{I_c}$ and $B = x^B\restriction_{I_c}$. Let $n$ be the length of the bit-strings $A$ and $B$.[10] We assume that $B$ is affected by noise such that $A \neq B$. Note that since the noisy quantum channel is i.i.d., each of the qubits is affected by noise with the same probability $p$, and thus (without loss of generality) we can assume that each bit of $B$ is affected by noise with probability $p$. Hence, we may model the situation by sending $A$ over a BSC with error-probability $p$. Note, however, that the BSC is not the physical channel, but refers to the theoretical channel model. In fact, the BSC represents a noisy *classical* channel. To make this more precise, note that Alice sends her $n$ qubits to Bob over the physical quantum channel. Only after Bob performs his measurements, he obtains classical information. However, if noise has occurred,it is possible that Bob's

[9] We will only describe the situation for $I_c$, as the situation for $I_{\bar{c}}$ is similar (we only need to replace $c$ by $\bar{c}$).

[10] As in the protocol, for $d \in \{0,1\}$, if the size $m$ of $I_d$ is smaller than $n$, we add $n-m$ zeros to $x^A\restriction_{I_d}$ to obtain an $n$-bit string, as this is according to [Sch10].

classical bit-string $B$ differs from the corresponding classical bit-string $A$ encoded by Alice, i.e./ that $x^B\restriction_{I_c} \neq x^A\restriction_{I_c}$. These differences can hence be represented by the BSC, even though in practice these classical bits were not transmitted over a classical channel.

As mentioned before, we use RS codes to correct the errors induced by the channel. We note that our method of information reconciliation is slightly different from the one described in the protocol from [Sch10]. In particular, [Sch10] considers the case that Alice sends the syndromes (which are of length $n-k$) of her two $n$-bit strings $x^A\restriction_{I_0}$ and $x^A\restriction_{I_1}$ to Bob. Instead, to keep the implementations relatively simpler, in our approach, Alice first encodes her two $n$-bit strings into codewords of length $m > n$, and then sends the last $m-n$ (redundant) bits of both of the codewords to Bob. Bob will attach the $m-n$ bits that belong to $x^A\restriction_{I_c}$ to his $n$-bit string $x^B\restriction_{I_c}$, and then use RS decoding techniques to obtain Alice's string $x^A\restriction_{I_c}$. Although our method might be slightly less efficient than the method provided in [Sch10], we note that the correctness and security of the protocol from [Sch10] still apply for our approach. Furthermore, we note that because in our approach $m-n$ additional bits are sent over the classical channel, privacy amplification must be adapted accordingly: the final length of $\ell$ of the output strings must be shortened by $m$ bits. Nevertheless, we will assume that when Alice and Bob agree on the values for $\ell$ and $n$ they already take into account this reduction in final string length.

The resulting protocol for information reconciliation is given below. All communication is achieved over a noiseless classical channel. We assume that both Alice and Bob are able to perform RS encoding and decoding and have agreed on a parameter $m$, which is the length of the encoded list and determines the maximal number of errors that can be corrected. Recall that given the length $n$ of Alice's and Bob's strings, using RS codes we can correct up to $\frac{m-n}{2}$ errors.

### Information reconciliation for 1-2 ROT

We assume that Alice holds two $n$-bit strings $x^A\restriction_{I_0}$ and $x^A\restriction_{I_1}$, and Bob holds an $n$-bit string $x^B\restriction_{I_c}$, where $c$ is his choice bit.

1. Alice encodes $x^A\restriction_{I_0}$ and $x^A\restriction_{I_1}$ using an RS encoder and sends the last $m-n$ bits of the encoded strings to Bob.

2. Bob attaches the $m-n$ bits corresponding to $x^A\restriction_{I_c}$ to his $n$-bit string $x^B\restriction_{I_c}$.

3. Bob uses an RS decoder to correct the errors in his resulting $m$-bit string, if possible.
   If the number of errors is greater than $\frac{m-n}{2}$, the protocol is aborted.

### Robust protocol for 1-2 OT

We will now provide the protocol for robust 1-2 ROT in the noisy-storage model on which our code is based. Protocol 4.2 is a slightly adapted from the protocol in [Sch10]. In particular, the protocol in [Sch10] takes into account the possibility that some of the qubits that are transmitted by Alice will eventually not be detected by Bob, which may result from imperfections in Bob's detection apparatus. However, in order to keep the implementation simple and understandable, we assume that all $n$ qubits arrive at Bob. Therefore, a few steps in the protocol from [Sch10] are omitted. We acknowledge, however, that it would be more realistic to assume that not all qubits are detected by Bob. Nevertheless, we emphasise that the correctness and security analysis from [Sch10] still hold in our case.

We use the notation from Section 3.3 and again assume that Alice and Bob agree on $\ell$ and a particular security error-probability, which determines $n$. Alice and Bob also agree on the value of m for information reconciliation.

### [Sch10] Robust 1-2 ROT$^l$

1. Alice randomly picks $x^A \in_R \{0,1\}^n$ and $y^A \in_R \{0,1\}^n$.

2. Bob randomly picks $y^B \in_R \{0,1\}^n$.

3. For each $i \in [n]$, Alice sends her bit $x_i^A$ encoded as $\mathrm{H}^{y_i^A}|x_i^A\rangle$ to Bob.
   If $y_i^B = 0$, Bob measures the $i$-th incoming qubit in the computational basis. Otherwise, he measures the qubit in the Hadamard basis.

Both parties wait time $\Delta t$.

4. Alice sends her basis string $y^A$ to Bob.

5. Bob forms the sets $I_c = \{i \in [n] \mid y_i^A = y_i^B\}$ and $I_{\bar{c}} = \{i \in [n] \mid y_i^A \neq y_i^B\}$, where $c$ is his choice bit.
   Bob sends $I_0$ and $I_1$ to Alice.

6. Alice performs privacy amplification: she picks two hash functions $f_0, f_1 \in_R \mathcal{F}$, where $\mathcal{F}$ is a class of two-universal hash functions $f : \{0,1\}^n \to \{0,1\}^l$. Alice sends sends $f_0, f_1$ to Bob.
   Alice performs her part of information reconciliation: she encodes $x^A\restriction_{I_0}$ and $x^A\restriction_{I_1}$ using an RS encoder and sends the last $m - n$ bits (say $r_0, r_1$) of the encoded strings to Bob.
   Alice outputs $s_0 = f_0(x^A\restriction_{I_0})$ and $s_1 = f_1(x^A\restriction_{I_1})$.

7. Bob performs his part of information reconciliation: Bob attaches $r_c$ to his own string and uses an RS decoder to correct the errors on his bit-string $x^B\restriction_{I_c}$. He obtains the corrected bit-string $x^B_{cor}$.
   Bob outputs $\hat{s_c} = f_c(x^B_{cor})$.

We emphasise that robust 1-2 ROT can be converted into robust 1-2 OT in a similar way as has been done for the idealised setting in Protocol 3.2. We therefore omit the resulting protocol.

# V. Quantum internet and SimulaQron

A quantum internet is a quantum communication network that allows for the exchange of quantum information over arbitrarily long distances [Cas18; DW18; DLH17]. One of the key aspects of a quantum internet is that it will allow for *quantum entanglement* between any two points in the world [CCB18; Cas18; DLH17; Kim08]. Quantum entanglement is a type of correlation between two quantum systems, which has two important features that can be used for several applications. The first is that two maximally entangled qubits are perfectly correlated: measuring one quantum state will immediately affect the outcome of measuring the other state as well [DLH17; CCB18; WEH18].[11] The second feature of quantum entanglement, complete privacy, is particularly interesting for security-related tasks: no more than two qubits can be maximally entangled, and hence no third party can participate in the entanglement [WEH18].

## Significance and applications of a quantum internet

It is expected that a quantum internet will be used alongside the classical internet rather than replacing it. [WEH18]. Therefore, we explain the significance of having quantum internet next to the classical internet. Moreover, we explain what the benefits may be of constructing a quantum internet rather than building a large-scale quantum computer. We conclude with some potential applications of a quantum internet.

### Comparison with the classical internet and the quantum computer

A quantum network may allow us to enlarge the state space used for quantum computation. In particular, with a quantum network we may not only surpass the capabilities of a classical network, but potentially also those of a single quantum processor.

Firstly, each node in a quantum network consists of a quantum processor, which has several advantages over a classical computer. In particular, a quantum computer allows us to exploit the laws of quantum mechanics—e.g. by making use of quantum entanglement—in order to tackle problems that are hard to solve, or unsolvable, with only classical means. Moreover, the state space of a quantum processor (i.e. the number of states on which it operates) scales exponentially with the number of qubits it entails as a consequence of the superposition principle [Kim08; CCB18]. Recall from Section 2 that a single qubit can exist in a superposition of two states, and therefore $n$ qubits can exist in a superposition of $2n$ states. Therefore, whereas a classical processor consisting of $n$ classical bits has a computational space of dimension $n$, a quantum processor consisting of $n$ qubits has a computational space of dimension $2^n$. Consequently, this larger state space of the quantum processors in the network may allow us to solve problems exponentially faster than would classically be possible [Kim08; WEH18].

However, the potential of a quantum internet is not only explained by the fact that its individual components—i.e. the quantum processors—may perform better than if they were substituted by classical processors, but also since *interconnect-*

[11] Note that we consider the Copenhagen interpretation of quantum mechanics.

*ing* the quantum processors with each other increases the overall computational power of the network. This feature is called "quantum connectivity" by [Kim08] and is explained as follows. Suppose that we have $k$ quantum processors, consisting of $n$ qubits each (i.e. each has computational space of dimension $2^n$). Then, if these quantum processors are connected via quantum channels, the overall computational space of the network is of dimension $2^{kn}$. In contrast, if these $k$ quantum processors (nodes) were connected via classical channels, then the computational space is only of dimension $k2^n$. In other words, because of its quantum connectivity, a quantum internet potentially allows for an improved performance than each of its quantum processors is capable of individually.

In particular, some researchers (e.g. [DLH17]) speculate that the realisation of a quantum internet is more feasible than building a large-scale quantum computer. The development of quantum computers places a maximum on the attainable size of individual quantum processors [Kim08], despite the fact that technological progress may increase this maximum over time. In contrast, a quantum internet is potentially able to exceed this maximum by interconnecting these individual processors, as previously explained. Moreover, many of the potential applications of a quantum network require only a quantum processor consisting of a single qubit, and thus not necessarily a large quantum computer. We may therefore consider quantum internet as a possible (temporary) solution for the challenges that prevail in building a large-scale quantum computer.

### Potential applications

The possibility of quantum communication over long distances opens many potential applications for a future quantum internet. One of the main expected applications of a quantum internet is quantum cryptography, and thus a quantum internet can be used to provide secure communication [DLH17]. For example, we could implement QKD [DLH17; WEH18][12] in order to establish a secure secret key between any two points in the network. Besides QKD, a quantum internet will also allow us to realise protocols for two-party cryptographic tasks [DW18; WEH18]. Another application of a quantum internet is distributed quantum computation [DLH17; Cac+18]: by linking different small-sized quantum computers via quantum channels [DLH17], we can scale up the computational space of each individual node, as previously explained. Therefore, a quantum internet consisting of multiple quantum processors may allow us to solve problems (possibly exponentially) faster than would be possible with only a single quantum processor. Other applications of a quantum internet include, but are not limited to, clock synchronisation [Kóm+14] and improving the achievable resolution of telescopes [GJC12].

## Towards the realisation of a large-scale quantum network

Having discussed the significance and some potential applications of a future quantum internet, we will now provide a high-level description of how a quantum internet could be constructed and analyse the feasibility of this construction. Before continuing, however, we first describe how quantum information is transmitted.

### Transmission of quantum information

The direct transmission of information in a classical or quantum communication network is affected by imperfections, i.e. by signal losses or noise. Classical information is often transmitted as electromagnetic waves over optical or free-space fibres [DLH17]. In the presence of noise or signal losses, the affected classical information can be recovered by reamplifying the signal at a so-called classical repeater [DLH17] or simply by resending the information [WEH18]. However, these classical solutions of amplification and repetition cannot be applied when the transmission of *quantum* information is affected by such imperfections. For direct transmission, quantum information is encoded as photons that are transmitted over optical or free-space fibres, similar to methods used in classical communication [DLH17]. However, if a transmitted photon is either lost or affected by noise, then the quantum information it carries is destroyed. More precisely, as a consequence of the no-cloning theorem [Die82; WZ82], which states that we cannot copy the quantum state of a qubit, we cannot re-

[12] The realisation of these cryptographic protocols obviously depends on the assumptions imposed by the protocol, e.g. that the adversary's quantum storage is noisy for the realisation of the 1-2 OT protocols in this paper.

cover the lost quantum information by sending it again. As a result, the occurrence of losses and dephasing errors during transmission only allow for sending photons over a few hundred kilometres [DLH17], and hence we need to find a different way to transmit quantum information over longer distances in order to establish a large-scale network.

Although quantum error correction may provide a method to overcome the photon losses and errors, and thus may still enable the direct transmission of photons over long distances [DLH17], further research in this direction is required before we can conclude whether this is an acceptable solution. More precisely, this approach is currently still challenging as it requires a relatively low error rate [DLH17]. Nevertheless, an alternative technique known as *quantum teleportation* was proposed, which enables the exchange of quantum information between two remote nodes [Ben+93]. Quantum teleportation is a method to transmit an unknown quantum state for which the sender and receiver are solely required to share two entangled qubits and to be connected by a classical communication channel. Hence, quantum teleportation does not require a quantum communication channel between the sender and the receiver at the moment of transmission. In particular, quantum teleportation achieves long-distance transmission of quantum information without physically sending the particle that stores it [Cac+18] and without violating the principles of quantum mechanics [NC10]. Since the quantum information is not transmitted directly (i.e. physically), quantum teleportation prevents the transmission from being affected by noise or photon losses.

Although we have explained that quantum teleportation is a crucial technique to establish a large-scale quantum network, teleporting quantum information between two remote nodes forces these two nodes to be entangled with each other. Once entangled, the distance between the two nodes can be arbitrarily long. However, *establishing* quantum entanglement between the nodes cannot be done over an arbitrarily long distance, since the entanglement distribution rate gradually decreases over distance [Cac+18]. In other words, before we can apply quantum teleportation over arbitrarily long distances we still need to find a way to provide entanglement over such a distance. In the next section, we explain that such distributed entanglement can be obtained by including intermediate nodes, called *quantum repeaters*, which apply a technique called *entanglement swapping*.

### Constructing a quantum internet

We have explained how one can transmit quantum information from one node to another, and will now describe how a quantum internet can be constructed.

The three concrete building blocks for a quantum internet are end nodes, photonic communication channels and quantum repeaters, which will be described below [WEH18] (see Figure 6 for a simplified illustration). The end nodes are quantum processors, which may vary from a single qubit to multiple qubits forming a large-scale quantum computer, depending on the task that must be performed.[13] At the end nodes, quantum states are generated, processed and stored [Kim08]. Moreover, these end nodes are connected to the network via a photonic channel—i.e. a quantum channel—to enable the transmission of (quantum) information and distribute entanglement across the network [Kim08]. Similar to the classical internet, this physical connection can be established via fibre-based channels (e.g. optical fibres), free-space channels, or a combination of these, provided that photon loss and decoherence (i.e. loss of quantum information stored in a qubit over time) is minimal [WEH18].

The third component of a quantum internet is a device that increases the distance through which the quantum information is transmitted: a quantum repeater. More precisely, a quantum repeater enables the distribution of entanglement over a long distance. Consequently, we can take advantage of this established entanglement to ‘teleport’ quantum information over this longer distance. As was explained before, quantum teleportation provides a solution to the imperfections that hinder the long-distance transmission of quantum information via photons. When the distance between the two end nodes is too large to establish the entanglement required to perform quantum teleportation, quantum repeaters are positioned along the communication channel as intermediate nodes to bridge this distance. In particular, quantum re-

[13] As will also be pointed out later on, Wehner, Elkouss and Hanson [WEH18] explain that not all applications of a quantum internet require a full-blown quantum internet; sometimes a slightly less developed “version” is already sufficient to perform a certain task.

peaters are placed in such a way that the distance to each end node is small enough to establish entanglement over the channel [WEH18]. After generating entanglement between itself and each of the end nodes, the quantum repeater implements entanglement swapping (see Figure 7): it teleports the entangled qubit that it shares with one of the end nodes to the other end node, such that the two end nodes eventually share an entangled qubit [WEH18]. In this way, entanglement is established over a distance that could be longer than the maximum distance possible via direct transmission [WEH18].

### Difficulties of establishing a quantum internet

From the above, it is clear that different issues arise when aiming to establish a quantum internet. These challenges are imposed by quantum mechanical features such as quantum entanglement and no-cloning [Cac+18]. Since these features do not have a classical counterpart, a considerable 'paradigm shift' is needed in order to counteract these barriers. An example of such a barrier is the occurrence of errors (e.g. due to decoherence or imperfect operations) when qubits interact with the environment [Cac+18]. Specifically, since quantum information cannot be copied, we are unable to apply classical error-correction methods that depend on information cloning, and thus a new solution is required [Cac+18]. Nevertheless, research is devoted to overcome the challenges posed by the development of a quantum internet [Kim08] and the realisation seems to be technologically possible, even though we are still at an early stage [DLH17]. Specifically, Wehner, Elkouss and Hanson [WEH18] even predict that a small-scale quantum internet will be established within a few years. Therefore, the current developments seem to provide an optimistic outlook for the establishment of a quantum internet.

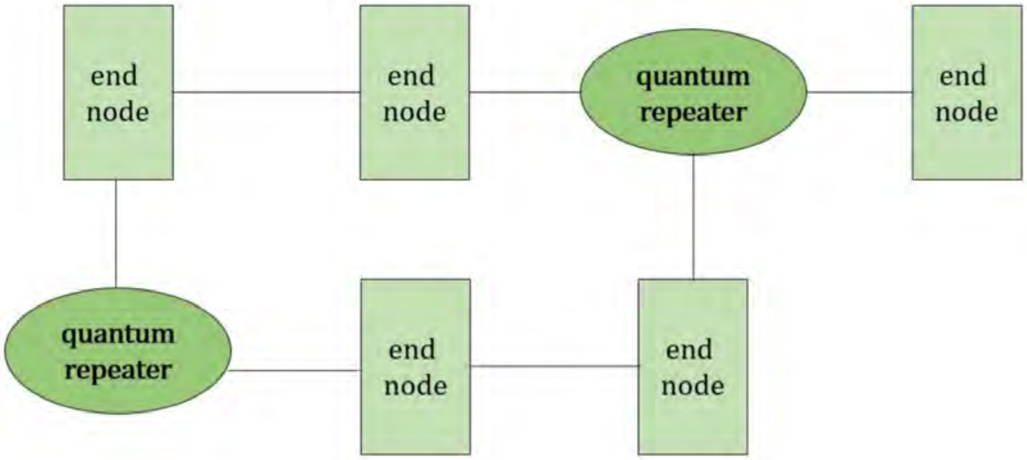


Figure 6: A simplification of the components of a quantum internet. The end nodes are the quantum processors where quantum states are generated, processed and stored. The quantum repeaters function as intermediate nodes, bridging the distances between the end nodes. The end nodes and quantum repeaters are connected via photonic channels (lines), over which quantum information is distributed.

### Secure two-party computation on a quantum internet

Fortunately, a fully developed world-wide quantum internet is not required to implement protocols for secure two-party computation. In particular, the process of establishing a quantum internet is divided into several developmental stages by [WEH18], including a description of the known applications of a quantum internet that can (already)

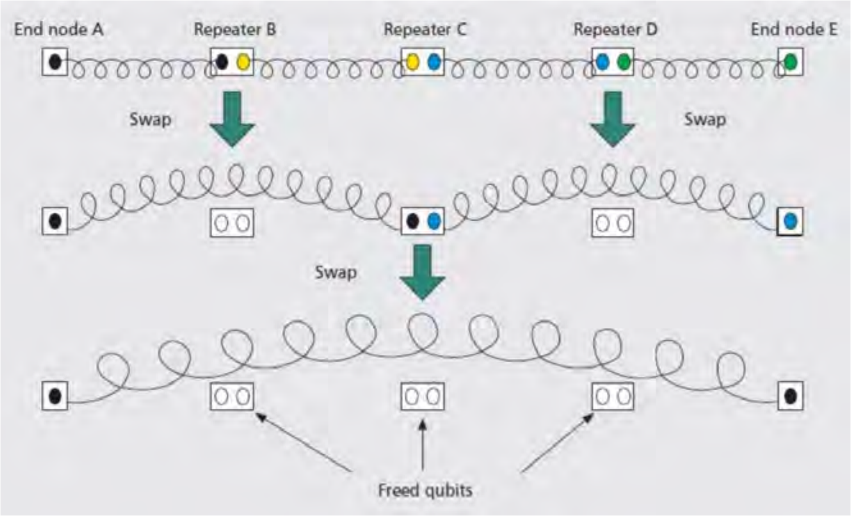


Figure 7: (Source: [MT13]) In order to increase the distance between two end nodes, one or more quantum repeaters are placed along the quantum channel that connects the end nodes. Here, three quantum repeaters are placed between end node A and end node E. For the end nodes and repeaters, entangled pairs are shared between adjacent nodes. First, repeater B teleports the qubit she shares with A to repeater C using her entanglement with repeater C. After this process of entanglement swapping, node A now shares an entangled pair with repeater C. In a similar way, entanglement swapping between repeater C, repeater D and end node E results in an entangled pair between repeater C and end node E. Then, once more entanglement swapping takes place to ensure that end node A and end node E share an entangled pair. All other qubits are freed.

be established at each stage. This work demonstrates that protocols for secure two-party computation can already, to some extent[14], be realised in the stage that allows for end-to-end transmission of qubits as well as the preparation and measurement of a quantum state at any node [WEH18]. Moreover, experimental results show that the essential elements required for this stage—including quantum repeaters—are within reach [WEH18]. Therefore, we may expect that the realisation of secure two-party computation is feasible in a near-future implementation of a quantum internet.

## Simulating quantum internet: SimulaQron

If a quantum internet will be deployed in the future, we will need adequate software for quantum-internet applications. For this reason, the quantum-internet simulator SimulaQron has been created [DW18]. SimulaQron has the purpose to serve as a framework in which software for quantum-internet applications can be written and debugged. To the best of our knowledge, SimulaQron is the only framework for developing software for quantum-internet applications currently available.[15]

### Working of SimulaQron

We will briefly describe the working of SimulaQron, which is visualised in Figure 8.[16] The virtual simulation network is established by classically connecting so-called *virtual nodes* with each other. A virtual node is labelled by a name (say Alice) and can be viewed as a server program running on a local classical computer that wants to connect to the network. The network can be simulated locally by running the different server programs all on one physical computer, but a distributed simulation is also possible (i.e. different virtual nodes are con-

[14] Nevertheless, the realisation of these protocols is only possible if the inaccuracies in transmission and measurement and the probability that the prepared quantum state is lost are below a certain level. It is still an open question, however, what the specific bounds are for these three parameters are [WEH18].

[15] Another quantum-network simulator called NetSquid is currently under development at QuTech in Delft, but not yet publicly available.

[16] For an extended description of the working of SimulaQron we refer the reader to [DW18].

nected to different computers). In other words, the virtual nodes corresponding to 'Alice Computer', 'Bob Computer', and 'Charlie Computer' in Figure 8 could either be run on a single computer, but also on two or three distinct classical computers. SimulaQron uses an existing quantum-hardware simulator to simulate the quantum processor on each virtual node, enabling a virtual node to simulate and manipulate qubits. By default, SimulaQron uses the stabilizer formalism[17] as quantum-hardware simulator, but it also supports ProjectQ [SHT18] and QuTip [JNN12].[18]

Besides simulating and manipulating qubits, a virtual node can also connect to other nodes in the network to enable classical and (simulated) quantum communication. This communication between virtual nodes is illustrated in Figure 8 by the lines between the different computers. The SimulaQron servers on the distinct virtual nodes allow for connecting the underlying simulated quantum hardware in order to establish (simulated) quantum communication and quantum entanglement, which is labelled "SimulaQron internal communication" in the figure. The classical-quantum-combiner (CQC) server in each node functions as a link between the SimulaQron back end and the level at which applications are written. In other words, applications on a particular computer (i.e. virtual node) communicate with the SimulaQron server via this CQC server. In addition, as quantum-network applications will often require classical communication as well, SimulaQron also allows for classical communication between distinct virtual nodes, labelled "application communication" in the figure.[19]

Figure 9 illustrates how we can program quantum-internet applications in SimulaQron. There are two ways to program the quantum network simulated by SimulaQron. The first method—called programming 'in native mode'—is by directly assessing the back end of SimulaQron using the Twisted [Twi] library for Python. However, the recommended way is via the provided classical-quantum-combiner (CQC) interface. The CQC interface could be perceived as an intermediate point between the application level and the SimulaQron back end, which allows the user to program at a higher level and hence facilitates the writing of code. Moreover, the CQC interface comes with both a Python and a C library that consist of useful methods to perform quantum operations in Python and C, respectively.

### Benefits of SimulaQron

Using SimulaQron has several advantages, as explained in [DW18]. First of all, SimulaQron provides software developers with a toolkit to write software that could later be put into practice on a real quantum internet with little or no adjustments. In particular, application development is independent of the underlying quantum hardware (as a result of the CQC interface) and is promoted by providing programming libraries for Python as well as C. Furthermore, the fact that it is written in Python facilitates a further extension of SimulaQron. Another benefit, which is of particular interest to our work, is that we may simulate noise in the channel by turning the setting `noisy-qubits` on. Hence, SimulaQron seems to provide the right tools in order to implement our protocols for 1-2 OT.

### Expected limitations

Nevertheless, SimulaQron also creates some drawbacks, which may become a limitation for our implementations. A first drawback of SimulaQron that is relevant to our implementations is that time is not modelled accurately, and hence we cannot simulate time-dependent noise [DW18]. It is therefore important to keep in mind that we cannot use SimulaQron to accurately evaluate the robustness of our protocols. Nonetheless, we are still able to simulate a noisy quantum channel, and hence can still implement our protocols. Another drawback of SimulaQron is that we cannot rely on its security for real usage. More precisely, since the entanglement created with SimulaQron is only a simulation of en-

[17] The stabilizer formalism is a specific approach to simulate so-called stabilizer circuits on a classical computer, Thesewhich are quantum circuits that are restricted to specific quantum operations. For example, see [Got97] for more information.

[18] The underlying back end can be changed to ProjectQ or QuTip in the settings via `simulaqron set backend`. However, one could also implement another existing simulator as long as it supports working with Python [DW18].

[19] More precisely, this classical communication can be established by opening a socket connection between the nodes. We are allowed to program such a client/server setup in Python, but the Python CQC library of SimulaQron also provides a built-in feature to establish classical communication via the methods `sendClassical` and `recvClassical` (see Section 6.1).

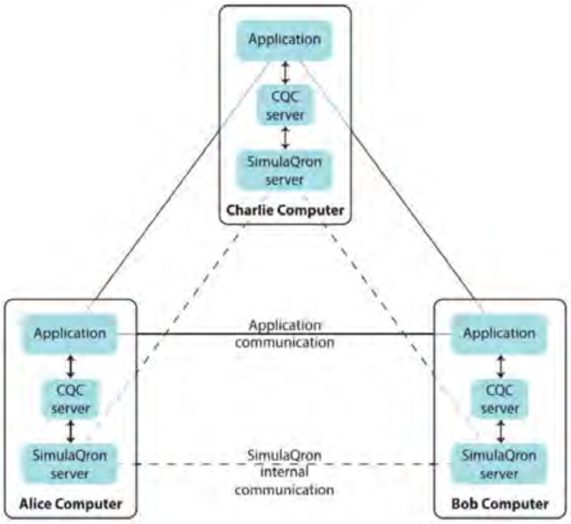


Figure 8: (Source: [DW18]) An illustration of the communication in a quantum network simulated by SimulaQron.

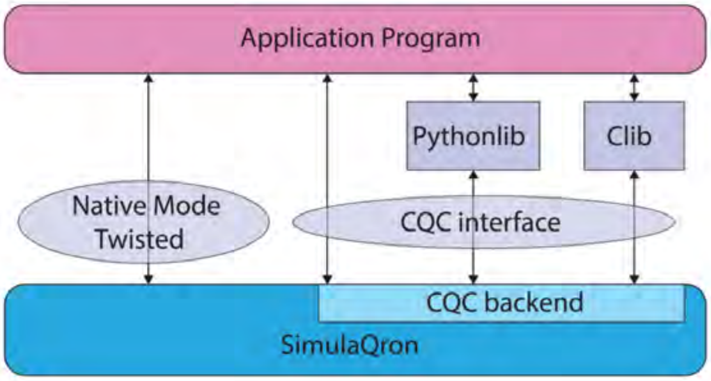


Figure 9: (Source: [DW18]) An illustration of programming a quantum network simulated by SimulaQron, which can be done either by directly assessing the SimulaQron back end using Twisted (left) or by running applications via the CQC interface (right).

tanglement, the security guarantees of real entanglement do not hold [DW18]. However, as we do not need to establish entanglement for our protocols, we do not expect this to become a limitation.

## VI. Implementation of the protocols

### Methods

In this section, we explain how we have implemented the protocols for 1-2 ROT (and 1-2 OT) in the noisy-storage model using SimulaQron and the Python CQC library. The version of SimulaQron we work with is version 3.0.3 and our code is written in Python [Pyt].

Before we continue, we provide our reasoning for the choice of the Python environment for our code. Firstly, SimulaQron is itself written in Python and comes with a Python CQC library that provides us with the right tools to implement our algorithms in the simulated quantum network. Therefore, our code corresponds to the examples provided in the SimulaQron documentation [Sim], making it understandable for current and future SimulaQron users. Secondly, Python is a well-known high-level programming language. We therefore expect that our code and implementations are clear to other users, which may allow for further optimisation and use.

### How to get started with SimulaQron

Section 5.3 explains how SimulaQron establishes a simulated quantum network. We will now explain in more detail how SimulaQron allows us to program our protocols and which tools of SimulaQron we have used for our implementations.

Before we can run our code we have to configure the simulated quantum network. Since we consider two parties for our protocols, the setup for our implementations only requires two virtual nodes: Alice and Bob. Although SimulaQron allows for a distributed simulation, we use the default configuration of SimulaQron in which the servers run in a centralised setting (i.e. localhost). In order to start the SimulaQron back end that consists of the virtual nodes we run the command `simulaqron start`, which will by default start a network with five nodes (labelled Alice, Bob, Charlie, David and Eve). We use the default back end (i.e. the stabilizer formalism), as this back end is the most efficient in the number of qubits [Sim]. We need to change two of the default settings for our implementations, which should be done before starting the back end. These two settings are `max-qubits`, the number of qubits allowed per quantum register in a virtual node (by default set to 20), and `noisy-qubits`, the option to apply noise to the qubits transmitted over the network (by default turned off).

Furthermore, for our implementations we make use of the functionalities provided by the Python CQC library SimulaQron. As explained in Section 5.3, this library allows us to program our protocols using the CQC interface without needing to access the SimulaQron back end directly. In order to use the library, we need to initialise a `CQCConnection` object, which takes as input the name of the node (say, Alice) to which it corresponds. A CQCConnection object enables the node to communicate with SimulaQron and with other virtual nodes. Several methods can be applied to a `CQCConnection` object, of which the following are of use for our work:

1. `sendQubit(q, name)`, which sends qubit q to node name;

2. `recvQubit()`, which receives a qubit sent to this node;

3. `sendClassical(name, msg)`, which opens a socket connection and sends `msg` (integer or list of integers) to name;

4. `recvClassical()`, which receives a classical message sent to this node and closes the socket connection.

The second tool from the Python CQC library that we use is the `qubit` object, which takes as input the corresponding `CQCConnection` (allowing for communication with SimulaQron) and is initialised to be in state $|0\rangle$. The useful methods for a `qubit` object are:

1. `X()`, which applies the X operator to the qubit

2. `H()`, which applies the Hadamard operator to the qubit

3. `measure()`, which measures the qubit and returns the outcome

For example, the following code corresponds to sending qubit q in state $|1\rangle$ to node 'Bob':

```
# initialise the connection
with CQCConnection("Alice") as Alice:
    # the code
    q = qubit(Alice)
    q.X()
    Alice.sendQubit(q, "Bob")
```

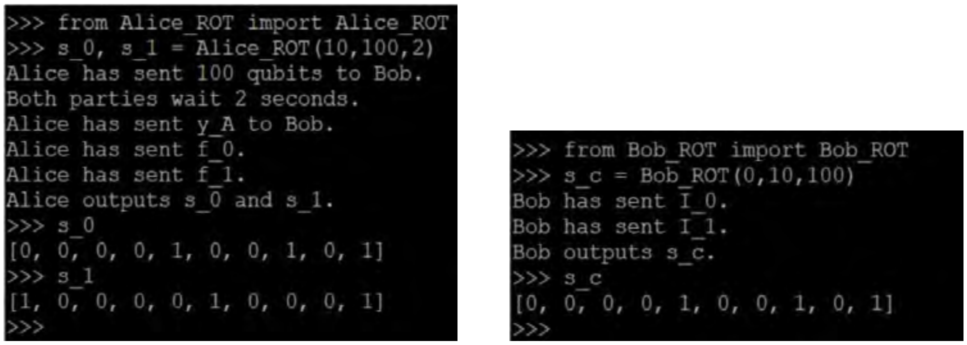
For a more explicit description of the possibilities of SimulaQron and the Python CQC library, we refer the reader to the documentation found in [Sim].

### Construction of the code

Our code is based on Protocol 3.1, 3.2, 4.1, and 4.2. Using the above methods provided by the Python CQC library, we can implement the required quantum and classical information in the protocol, as well as the manipulation of the qubits (i.e. the encoding of Alice's random string xA as qubits to be sent to Bob). In order to facilitate the computations in the code, Alice's and Bob's strings (`s_0`, `s_1`, `m_0`, and `m_1`) are programmed as Python `list` objects. Furthermore, during the protocol we print several messages to indicate that communication between Alice and Bob has occurred. For the privacy amplification step we use random binary matrices to serve as two-universal hash functions.

### Additional methods for robust 1-2 OT

As explained in Section 4, in a realistic setting we may expect that the quantum channel is



```
>>> from Alice_ROT import Alice_ROT
>>> s_0, s_1 = Alice_ROT(10,100,2)
Alice has sent 100 qubits to Bob.
Both parties wait 2 seconds.
Alice has sent y_A to Bob.
Alice has sent f_0.
Alice has sent f_1.
Alice outputs s_0 and s_1.
>>> s_0
[0, 0, 0, 0, 1, 0, 0, 1, 0, 1]
>>> s_1
[1, 0, 0, 0, 0, 1, 0, 0, 0, 1]
>>>
```

(a) Alice's terminal

```
>>> from Bob_ROT import Bob_ROT
>>> s_c = Bob_ROT(0,10,100)
Bob has sent I_0.
Bob has sent I_1.
Bob outputs s_c.
>>> s_c
[0, 0, 0, 0, 1, 0, 0, 1, 0, 1]
>>>
```

(b) Bob's terminal

Figure 10: 1-2 ROT for Alice and Bob. Alice receives two lists, s_0 and s_1. Bob holds choice bit 0 and receives Alice's output list s_0.

noisy. In order to implement Protocol 4.2 for robust 1-2 ROT we therefore need to find a way to deal with the errors induced by the channel. As explained, we have decided to work with Reed-Solomon (RS) codes and use the `reedsolo` library [Fil]. We are aware that there are more efficient encoding and decoding techniques available, such as sum-product decoding for low-parity-density-check (LDPC) codes as introduced by [Gal62]. However, there is to the best of our knowledge no suitable open-source implementation of such decoding algorithms that could be easily extended to our setting. Moreover, our implementations are currently not designed with the purpose to be as efficient as possible, and we do not intend to work with very large values for the length of the encoded word. Therefore, we do not expect that using the slightly less efficient Reed-Solomon coding techniques will introduce significant problems.

Furthermore, in order to simulate the noise in the quantum channel, we turn the setting `noisy-qubits` on. However, as also mentioned in [DW18], SimulaQron currently does not allow for an accurate simulation of noise. Although the setting `noisy-qubits` induces some noise in the channel, and the level of noise can be somewhat monitored via the setting `t1`, the resulting noise turned out to be unrealistic: the error-probability[20] approaches $\frac{1}{2}$ as the number of qubits increases, which implies that the received information is completely random. Therefore, we have made a change to the underlying SimulaQron code by manually setting the error-probability to be $\frac{1}{10}$, as this would be more plausible in a real setting.

## Results

We have implemented both the idealised and the robust protocols for 1-2 ROT and 1-2 OT in the noisy-storage model using SimulaQron. The code is written in Python and available on GitHub (https://github.com/lengelberts/simulaqron-qc). The source code can also be found in Appendix B. We emphasise that in order to run our code both `simulaqron` and `cqc` need to be installed as specified in the SimulaQron documentation [Sim].

### Implementation of the idealised protocols

We have implemented Protocols 3.1 and 3.2 for 1-2 ROT and 1-2 OT, respectively, in the noisy-storage model using SimulaQron. We have tested our code both in the setting of a noiseless quantum channel and in the setting that Alice and Bob are connected by a noisy quantum channel, which is simulated by turning the option `noisy-qubits` on and manually setting the error-probability to be $p = \frac{1}{10}$. In the presence of noise, running our implementation of the idealised protocol results in an incorrect output for Bob. So the idealised protocol breaks down in a noisy setting. We provide a few examples of running our code. [21]

**Example 1.** The first example is our implemen-

[20] Note that this error-probability is the probability parameter for the BSC model, as explained in Section 4.1.

[21] Note that our examples have the purpose to show the working of our implementations. In reality the number n of transmitted qubits is very large, much larger than we will consider for our examples.

tation of Protocol 3.1 for 1-2 ROT in the idealised setting. 1-2 ROT is realised by the functions called `Alice_ROT(l, n, waiting_time)` and `Bob_ROT(c, l, n)`. Here, the length of the output lists is `l = 10` and Bob holds choice bit `c = 0`. We also set `n = 100` and the waiting time in seconds between Step 2 and Step 3 is `waiting_time = 2`. Before starting the SimulaQron back, we set the maximal number of qubits to 500, as we want to transmit `n = 100` qubits (i.e. any other number greater or equal to 100 would also work). We therefore run the following commands:

```
$ simulaqron set max-qubits 500
$ simulaqron start
```

We run the code in two separate Python processes,[22] one for Alice and one for Bob.
In Alice's terminal we run:

```
>>> Alice_ROT(l=10, n=100, waiting_time=2)
```

In Bob's terminal we run:

```
>>> Bob_ROT(c=0, l=10, n=100)
```

The resulting prints and outputs are illustrated in Figure 10. As the figure illustrates, Bob has indeed received Alice's output list `s_0`, corresponding to his choice bit `c = 0`.

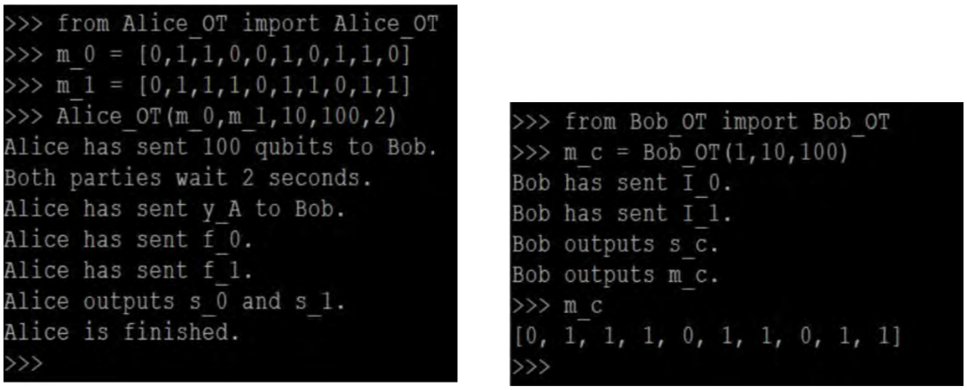
**Example 2.** In our second example, we show our implementation of Protocol 3.2 for 1-2 OT in the idealised setting.
For 1-2 OT, we wrote two functions called `Alice_OT(m_0, m_1, l, n, waiting_time)` and `Bob_OT(c, l, n)`. Again, we set `l = 10`, `n = 100`, and `waiting_time = 2`. However, Bob now holds choice bit `c = 1` and, since we will now run 1-2 OT, Alice holds two 10-bit lists `m_0 = [0,1,1,0,0,1,0,1,1,0]` and `m_1 = [0,1,1,1,0,1,1,0,1,1]`. We use the same settings as in Example **??**.
In Alice's terminal we now run:

```
>>> Alice_OT(m_0, m_1, l=10, n=100, waiting_time=2)
```

In the second Bob's terminal we run:

```
>>> Bob_OT(c=1, l=10, n=100)
```

The resulting prints and outputs are illustrated in Figure 11. As expected, Bob receives Alice's input list `m_1`, corresponding to his choice bit `c = 1`.



```
>>> from Alice_OT import Alice_OT
>>> m_0 = [0,1,1,0,0,1,0,1,1,0]
>>> m_1 = [0,1,1,1,0,1,1,0,1,1]
>>> Alice_OT(m_0,m_1,10,100,2)
Alice has sent 100 qubits to Bob.
Both parties wait 2 seconds.
Alice has sent y_A to Bob.
Alice has sent f_0.
Alice has sent f_1.
Alice outputs s_0 and s_1.
Alice is finished.
>>>
```

(a) Alice's terminal

```
>>> from Bob_OT import Bob_OT
>>> m_c = Bob_OT(1,10,100)
Bob has sent I_0.
Bob has sent I_1.
Bob outputs s_c.
Bob outputs m_c.
>>> m_c
[0, 1, 1, 1, 0, 1, 1, 0, 1, 1]
>>>
```

(b) Bob's terminal

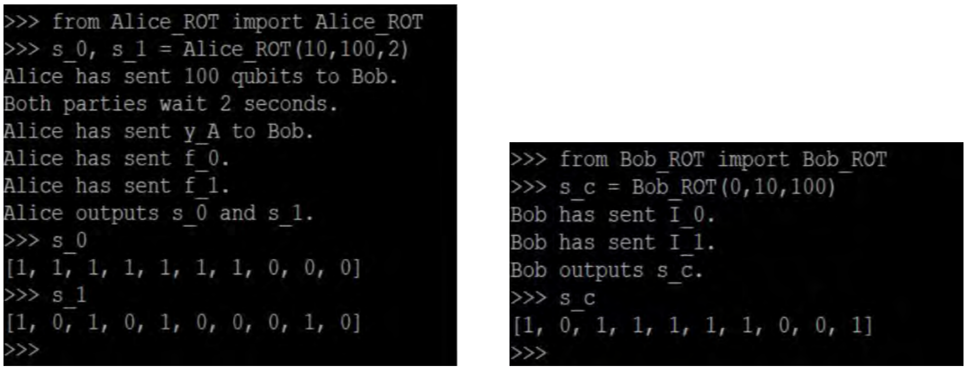
Figure 11: 1-2 OT for Alice and Bob. Alice inputs m_0 and m_1. Bob holds choice bit 1 and receives m_1.

**Example 3.** In our third example, we show what happens when we run 1-2 ROT in a noisy-setting. Hence, we use the functions `Alice_ROT` and `Bob_ROT`. We therefore first have to stop the SimulaQron back end, change the settings, and then start the back end again. We do this by typing the following commands:

```
$ simulaqron stop
$ simulaqron set noisy-qubits on
$ simulaqron start
```

Note that this will automatically set the value for `t1` to be 1. We use the same inputs as for our first example. That is, `c = 0`, `l = 10`, `n = 100`, and `waiting_time = 2`.

[22] Our code needs to be run in two separate Python processes in order to enable Alice and Bob to exchange classical information via the `sendClassical` and `recvClassical` commands, because the `sendClassical` method will wait until a socket is set up to a remote node and will wait forever if such a socket is never set up. Nevertheless, starting the SimulaQron back end and changing the settings can be done from within one of the two terminals.



```
>>> from Alice_ROT import Alice_ROT
>>> s_0, s_1 = Alice_ROT(10,100,2)
Alice has sent 100 qubits to Bob.
Both parties wait 2 seconds.
Alice has sent y_A to Bob.
Alice has sent f_0.
Alice has sent f_1.
Alice outputs s_0 and s_1.
>>> s_0
[1, 1, 1, 1, 1, 1, 1, 0, 0, 0]
>>> s_1
[1, 0, 1, 0, 1, 0, 0, 0, 1, 0]
>>>
```

(a) Alice's terminal

```
>>> from Bob_ROT import Bob_ROT
>>> s_c = Bob_ROT(0,10,100)
Bob has sent I_0.
Bob has sent I_1.
Bob outputs s_c.
>>> s_c
[1, 0, 1, 1, 1, 1, 1, 0, 0, 1]
>>>
```

(b) Bob's terminal

Figure 12: 1-2 ROT for Alice and Bob in a noisy setting. Alice receives two lists, s_0 and s_1. Bob holds choice bit 0, but does not receive the same output as Alice's output list s_0.

From Python in Alice's terminal, we run:

```
>>> Alice_ROT(l=10, n=100, waiting_time=2)
```

In the second terminal we run:

```
>>> Bob_ROT(c=0, l=10, n=100)
```

The resulting prints and outputs are illustrated in Figure 12. However, the list received by Bob does no longer correspond to Alice's output list `s_0`. The present noise has affected the output.

### Implementation of the robust protocol

In addition, we have considered the case for the robust protocol for 1-2 OT in the presence of noise. We have therefore investigated the noise model that is used in SimulaQron (when turning the option `noisy-qubits` on). As was previously briefly outlined, we may manipulate the error-probability via the setting `t1`, which represents the coherence time of the qubits: the lower the value, the more noise is added to the channel [Sim]. Nevertheless, as we mentioned before, the noise in the channel is unrealistic, regardless of how we tune `t1`, and thus we decided to fix manually the error-probability to be $p = \frac{1}{10}$.

However, regardless of whether we allow noise in the channel or not, the implementation of the robust protocol for 1-2 ROT turned out to be currently infeasible. More precisely, we wrote the two functions `Alice_robust_ROT` and `Bob_robust_ROT` according to Protocol 4.2. Yet, when we run the code, the Python terminals no longer respond, which occurs more or less halfway through the information reconciliation part of the code. We have experimented with our code in order to find out what may cause the problem. In particular, we have provided a separate code for the part of robust 1-2 ROT during which information reconciliation (i.e. according to Protocol 4.1) is applied. The example below will illustrate that our code for information reconciliation results in the expected outcome and is a working code. However, we note that we obtain the same problem as when running our implementation of Protocol 4.2 if Alice runs her code before Bob (i.e. before we have called the function for Bob). More precisely, the Python terminal again stops responding. Nevertheless, as the code returns the desired outcomes in the case that Bob first runs his code, we expect that the problem is not caused by our use of the `reedsolo` library, but by something else.

From further experimentation with the code, we speculate that the problem is caused by the way classical communication is achieved: via the methods `sendClassical` and `recvClassical`. We recall from Section 6.1 that each time `sendClassical` is called by (say) Alice, a socket connection is opened to Bob, which is only closed after Bob has received his message by `recvClassical`. In fact, note that the information reconciliation protocol requires Alice to send two more classical strings to Bob (i.e. the two $m-n$ strings $r_0$ and $r_1$). Therefore, we presume that the problem is caused by an imperfect opening or closing of the socket connection between Alice and Bob during this additional classical information. However, we emphasise the need to investigate further this encountered problem in order to eventually realise our robust 1-2 ROT implementation.

Despite the fact that running the code currently does not work, we still provide the code of our implementation of the robust protocol for 1-2 ROT in Appendix B. However, as previously explained, we have been able to provide a running code for Protocol 4.1 which establishes the information reconciliation part needed for robust 1-2 ROT. The two implemented functions are called Alice_reconciliation and Bob_reconciliation. We provide an example of usage below. Recall that for the setting of 1-2 ROT in a *noisy* quantum channel, when information reconciliation is applied, Alice has two strings $x_0$ and $x_1$, and Bob holds a string $\tilde{x_c}$ for his choice bit $c$, say 0. In particular, we may assume that the noise in the quantum channel entails that $\tilde{x_c} \neq x_0$. Hence, the task of information reconciliation is to achieve $x_c$ for $c = 0$.

**Example 4.** We illustrate our implementation of Protocol 4.1. Our implementation only requires classical information to be transmitted, and hence we do not need to change any of the default settings and can start SimulaQron immediately:

```
$ simulaqron start
```

Alice holds two binary lists of length $10$: namely x_0 = [0,1,1,1,0,0,1,1,0,1] and x_1 = [0,1,0,0,0,1,0,0,1,0]. Bob holds choice bit c = 0 and x_c = [0,1,1,0,0,0,1,0,0,1]. Note that the lists x_0 and x_c differ in two elements, since we assume that the quantum channel is noisy. Moreover, Alice and Bob agree on m = 16. Recall that RS codes can correct up to $\frac{m-n}{2} = 3$ errors, i.e. we may expect that the reedsolo codes from **?** will be able to correct the two errors on Bob's x_c. Indeed, that is what we see below.
In Alice's terminal we run:

```
>>> Alice_reconciliation(x_0, x_1, m=16, n=10)
```

In Bob's terminal we run:

```
>>> Bob_reconciliation(c=0, x_c, m=16, n=10)
```

The resulting prints and outputs are illustrated in Figure 13. As the figure illustrates, Bob returns Alice's input list x_0, corresponding to his choice bit c = 0.

## Discussion and implications

In the previous section we have provided the results of our implementations and some examples. We will now discuss what these results entail in greater detail.

### Potential of SimulaQron

First of all, our implementations elucidate the potential and some current limitations of SimulaQron. SimulaQron turns out to be an accessible platform for users to explore the possibilities of a potential future quantum network. SimulaQron is written in Python and provides a Python CQC library which consists of an extensive number of useful commands that facilitate the user to simulate quantum communication and to apply quantum operations. In particular, since a version of this CQC interface is intended to be available on the expected quantum network in the Netherlands in 2020 [DW18], software written for SimulaQron could possibly be extended for use on this real quantum network.

In addition, recall that SimulaQron has been designed with the goal to provide a developmental framework for writing and testing software for quantum-internet applications, quantum cryptography among these. Despite some limitations, our results demonstrate the feasibility of implementing 1-2 OT in SimulaQron, and hence the possibility of extending it to implementing any other secure two-party computation protocol. Furthermore, since the quantum part of the considered protocols is similar to that of the BB84 protocol [BB84] for QKD, our implementations also suggest the possibility of a QKD implementation in SimulaQron. Hence, our results indicate the potential of using SimulaQron as environment for implementing quantum-cryptographic protocols. In particular, our results show that SimulaQron could be used for quantum-cryptographic implementations as long as a realistic time-dependent noise model is not desired for testing and the user (i.e. the writer of the code) has sufficient programming experience. However, note that the security of SimulaQron itself is not comparable to the security of a real quantum internet. Therefore, we emphasise that SimulaQron is intended as a framework to develop software for quantum-internet applications such as quantum cryptography, and not actually to use this software.

Nevertheless, we encountered two main problems for our implementations which explain some limitations with respect to SimulaQron. Firstly, the methods provided by the Python CQC library (sendClassical and recvClassical) to enable classical communication have very limited possibilities. These methods merely allow us to transmit (lists of) integers and not, for example, arrays or Python str objects. In addition, we speculate that the current method for establishing a socket connection between any two nodes for classical communication is causing a problem and explains why our implementation of 1-2 ROT did not fully succeed. Although we are aware that SimulaQron also allows for the implementation of one's own socket connection [DW18], we consider a built-in feature for classical communication essential, since many quantum-cryptographic applications will require the use of classical communication [DW18]. For these reasons, we recommend further updating the possibilities of the Python CQC library to allow for more realistic classical communication.

A second limitation was that noise is not accurately modelled by SimulaQron, as is also explained in [DW18]. Although our results show that the noise model allows us to demonstrate that protocols may break down in the case of a noisy quantum channel, the level of noise it achieves is unrealistic, as previously explained. For this reason, it becomes difficult to demonstrate the working of protocols in a noisy setting. Nevertheless, by manually setting the error-probability to a certain (time-independent) level, we were able to get a more realistic level of noise, and we expect that this approach may be sufficient for future users as well.

Furthermore, we note that SimulaQron is still at a developmental stage: in the period of this work SimulaQron has been updated from version 1.3 to version 3.0.3. The main difficulties we encountered when using SimulaQron generally occurred during the installation of SimulaQron and when updating to the newer versions. In particular, we encountered problems when installing SimulaQron in Windows. Moreover, one needs to be careful when using SimulaQron as even minor errors such as typos could result in SimulaQron malfunctioning. Although these difficulties have provided significant obstacles during the process of writing our code, the updated newer versions of SimulaQron also allowed for more consistent and clear documentation. In addition, the frequent updates of SimulaQron forced us to be very aware of any new changes in SimulaQron, which resulted in a deeper understanding of the working of SimulaQron and its possibilities.

### Quantum internet and quantum cryptography

Our implementations provide an illustration of performing the task of 1-2 OT over a potential quantum internet. Secure two-party computation is one of the potential applications of a quantum internet [WEH18], and this thesis shows how such cryptographic applications may be realised in a quantum network. Moreover, we expect that our implementations might be of use in a future quantum internet: the protocols on which our code is based are designed in a relatively simple way and our code is written in Python, a well-known, high-level language. Nevertheless, we speculate whether we would really need a quantum internet to implement secure two-party computation. In fact, secure two-party computation is already shown to be possible with QKD hardware (e.g. see [Erv+14]). Moreover, it is questionable whether we really need a large-scale network in order to implement the considered protocols in practice, as it is reasonable to consider that secure two-party computation is performed at a short distance [Ben+92b].

In particular, we have not only showed how the idealised setting of 1-2 OT (for the noisy-storage model) can be implemented in a future quantum internet, but have also considered a more practical setting, namely when the quantum communication is affected by noise. Although we have not been able to run our code for robust 1-2 ROT in SimulaQron, we have demonstrated the need for such protocols, since the idealised protocols break down in a noisy setting. Moreover, we may expect that it will eventually be feasible to implement (a version of) our code for robust 1-2 ROT in SimulaQron as we have provided an explicit description of information reconciliation in the context of 1-2 OT and a working code to establish it.

### Other implications of our work

One of the benefits of our work is that our implementations are all publicly available on GitHub. Also, the decoding algorithm[23] that we have chosen to use for the implementation of the robust protocol is publicly accessible. This is in contrast to the currently existing implementations of 1-2 OT (namely those implemented using QKD hard-

[23] However, we note that our code is most likely not optimal, since most techniques previously used are not available for public use but are probably "better" in terms of efficiency. Therefore, we recommend further improvements be made to our code.

ware), for which the used code is licensed and (to our knowledge) not publicly available. Our work is more easily accessible than other implementations of 1-2 OT, and could therefore be used for future research in this direction and to further improve the code.

Moreover, as our implementations are simple, clear, and publicly accessible, our work could be used for learning purposes. More precisely, since we have explicitly described the steps in our work and since our code is properly documented, it may be a good source for newcomers in the field as it explains both the working of SimulaQron (and hence a future quantum internet) and of secure two-party computation in the setting of a noisy-quantum memory. Previous work, such as in [Sch10] and [KWW12], provides a complete description of 1-2 OT in the noisy-storage model, yet some steps are not included. Although these steps are trivial to experts in the field, our work may be a valuable clarification for newcomers and outsiders. In addition, our code could potentially be used to provide more examples in the SimulaQron documentation, as the documentation currently consists of a limited number of examples.

## VII. Conclusion

The two primary aims of this study were (i) to ascertain the suitability of the protocols of [Sch10] for quantum-internet applications and (ii) to investigate the extent to which SimulaQron is an adequate environment for implementing and analysing quantum-cryptographic protocols.

To the best of our knowledge we have provided the first implementations of quantum-cryptographic protocols in SimulaQron. As a result, this thesis elucidates the potential of SimulaQron and some of its limitations. In particular, our implementations show that SimulaQron may be a suitable environment for implementing quantum-cryptographic protocols, and indicate some issues that may be overcome in the future. In this way we contribute to a further development of SimulaQron, but also to the current development of a quantum internet. More precisely, we have illustrated how quantum-cryptography may be used on a future quantum internet. The analysis in our work may therefore be significant as quantum cryptography is one of the expected applications of a future quantum internet. On the other hand, we contributed to the field of quantum cryptography by providing a quantum-internet simulation of protocols for 1-2 OT. Additionally, our work shows that the protocols provided in [Sch10] are indeed adequate for quantum-internet applications and are feasible to implement.

### Future research

At the same time, our work reveals several potential areas for future research.

Firstly, we emphasise that our code is not currently designed to be as efficient as possible and could be further optimised. In particular, future research may be devoted to improving the decoding part in our implementations of robust 1-2 OT. In addition, it would be valuable to further investigate why our code for robust 1-2 OT currently does not work, which may become beneficial for the development SimulaQron as well. Moreover, as previously explained, the current built-in client/server setup for classical communication in SimulaQron is quite restrictive and inefficient, and hence we would advise that this feature be improved in the future.

Furthermore, it may be noteworthy to consider the simulation of possible adversarial attacks. Although the security of the implemented protocols is already proven for the given parameters, such an imitation of possible attacks may further clarify the setting of secure two-party computation. This could prove useful for a better understanding of these protocols, but may also explain to outsiders of the field the significance and relevance of such quantum-cryptographic protocols.

Another significant research direction may be to compare our implementations with existing implementations of 1-2 OT using QKD hardware. Although our quantum-cryptographic implementations are, to the best of our knowledge, the first ones in the setting of a quantum internet, there are already experimental implementations of 1-2 OT using QKD hardware (e.g. see [Erv+14]). Therefore, it would be an interesting direction of future research to compare the performance and feasibility of the implementations in both settings. However, it should be noted that—as previously argued—SimulaQron does not allow for a realistic comparison of time performance, and therefore we recommend using another simulator (potentially

NetSquid) to continue in this direction.

Lastly, SimulaQron is, to date and to our knowledge, the only quantum-internet simulator. However, as another simulation framework, NetSquid [Net], is under development, it would be interesting to compare SimulaQron's working, qualities and limitations with those of NetSquid. In particular, we have explained that although our results show that SimulaQron provides a suitable developmental framework for quantum-internet applications, it is not ideal. Therefore, it may be interesting to investigate whether NetSquid overcomes the limitations of SimulaQron.

# Appendix A

## Probability theory

We define some basic notions of probability theory, using the definitions from [DS17].

**Definition 9. (Probability measure)**
A probability measure $\mathbb{P}$ is a function $\mathbb{P} : \Omega \to \mathbb{R}_{\geq 0}$ such that

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1.$$

**Definition 10. (Probability space)**
A *(discrete) probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega$ is a non-empty sample space, $\mathcal{F}$ is an event space, and $\mathbb{P}$ is a probability measure.

**Definition 11. (Random variable)**
A *discrete random variable* $X$ on a discrete probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathcal{X}$, where $\mathcal{X}$ is a (discrete) set.

**Definition 12. (Probability distribution)**
Let $X$ be a random variable on the set $\mathcal{X}$. The *probability distribution* of $X$ is a function $P_X : \mathcal{X} \to [0,1]$ defined as

$$P_X(x) := \mathbb{P}[X = x],$$

where $X = x$ denotes the event $\{\omega \in \Omega | X(\omega) = x\}$. Note that $P_X$ is often called the *marginal* probability distribution of $X$.

**Definition 13. (Joint probability distribution)**
Let $X$ and $Y$ be two random variables defined on the same probability space, with ranges $\mathcal{X}$ and $\mathcal{Y}$, respectively. The pair $XY$ is a random variable and has probability distribution $P_{XY} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ given by

$$P_{XY}(x,y) := \mathbb{P}[X = x, y = \mathcal{Y}].$$

**Definition 14. (Conditional probability distribution)**
If $\mathcal{A}$ is an even such that $P[\mathcal{A}] > 0$, then we define the conditional probability distribution of $X$ given $\mathcal{A}$ by

$$P_{X|\mathcal{A}}(x) := \frac{P[X = x, \mathcal{A}]}{P[\mathcal{A}]}.$$

If $Y$ is another random variable and $P_Y(y) > 0$, then we write

$$P_{X|Y}(x|y) := P_{X|Y=y}(x) = \frac{P_{XY}(x,y)}{P_Y(y)}$$

for the conditional distribution of $X$ given $Y = y$.

# Appendix B

We provide the source code for our implementations. The code is also available on GitHub, click here.

## Works Cited

[ALA18] Ebrahim Ardeshir-Larijani and Farhad Arbab. “Reo coordination model for simulation of quantum internet software”. In: *Software Technologies: Applications and Foundations. STAF 2018.* Ed. by Manuel Mazzara, Iulian Ober, and Gwen Salaün. Vol. 11176. Lecture Notes in Computer Science. Springer, Cham, 2018, pp. 311–319. isbn: 9783030047702.

[BB84] C. H. Bennett and G. Brassard. “Quantum cryptography: Public key distribution and coin tossing”. In: *Proceedings of IEEE International Conference on Computers, Systems, and Signal Processing.* Bangalore, 1984, pp. 175–179.

[BBR88] Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert. “Privacy Amplification by Public Discussion”. In: *SIAM Journal on Computing* 17.2 (1988), pp. 210– 229. doi: 10.1137/0217014.

[Ben+92a] Charles H. Bennett et al. “Experimental quantum cryptography”. In: *Journal of Cryptology* 5.1 (1992), pp. 3–28. issn: 1432-1378. doi: 10.1007/BF00191318. url: https://doi.org/10.1007/BF00191318.

[Ben+92b] Charles H. Bennett et al. “Practical quantum oblivious transfer”. In: vol. 576.

Springer Verlag, 1992, pp. 351–366. isbn: 9783540551881.

[Ben+93] Charles H. Bennett et al. “Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels”. eng. In: *Physical review letters* 70.13 (1993). issn: 1079-7114.

[BS94] Gilles Brassard and Louis Salvail. “Secret-key reconciliation by public discussion”. In: vol. 765. Springer Verlag, 1994, pp. 410–423. isbn: 9783540576006.

[BCS12] Harry Buhrman, Matthias Christandl, and Christian Schaffner. “Complete insecurity of quantum protocols for classical two-party computation”. In: *Physical Review Letters* 109.16 (2012), p. 160501. doi: 10.1103/PhysRevLett.109.160501. arXiv: 1201.0849v2.

[Cac+18] Angela Sara Cacciapuoti et al. “Quantum Internet: Networking Challenges in Distributed Quantum Computing”. In: (Oct. 19, 2018). arXiv: 1810.08421v2 [quant-ph].

[CCB18] Marcello Caleffi, Angela Sara Cacciapuoti, and Giuseppe Bianchi. “Quantum internet: from communication to distributed computing!” In: (2018). arXiv: 1805. 04360v1.

[CW79] J.Lawrence Carter and Mark N. Wegman. “Universal classes of hash functions”. In: *Journal of Computer and System Sciences* 18.2 (1979), pp. 143–154. doi: 10.1016/0022-0000(79)90044-8.

[Cas18] Davide Castelvecchi. “The quantum internet has arrived (and it hasn’t)”. In: *Nature* 554 (2018), pp. 289–292. doi: 10.1038/d41586-018-01835-3.

[DW18] Axel Dahlberg and Stephanie Wehner. “SimulaQron—a simulator for developing quantum internet software”. In: *Quantum Science and Technology* 4.1 (2018), p. 015001. doi: 10.1088/2058-9565/aad56e.

[Dam+09] Ivan Damgaard et al. “Improving the Security of Quantum Protocols via Commitand-Open”. In: *Advances in Cryptology - CRYPTO 2009*, LNCS 5677, pages 408-427 (Feb. 23, 2009). arXiv: 0902.3918v4 [quant-ph].

[Dam+05] Ivan B. Damgård et al. “Cryptography in the bounded quantum-storage model”. In: *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science — FOCS 2005* (2005), pp. 449–458. arXiv: quant-ph/0508222v2.

[Dam+07] Ivan B. Damgård et al. “A tight high-order entropic quantum uncertainty relation With applications”. In: *Advances in Cryptology — CRYPTO 2007*. Vol. 4622. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2007, pp. 360– 378. arXiv: quant-ph/0612014v2.

[Die82] Dennis Dieks. “Communication by EPR devices”. In: *Physics Letters* A 92.6 (1982), pp. 271–272. doi: 10.1016/0375-9601(82)90084-6.

[DS17] Yfke Dulek and Christian Schaffner. *Information Theory*. Lecture Notes. University of Amsterdam, Master of Logic. 2017.

[DLH17] Wolfgang Dür, Raphael Lamprecht, and Stefan Heusler. “Towards a quantum internet”. In: *European Journal of Physics* 38.4 (2017), p. 043001. doi: 10.1088/ 1361-6404/aa6df7.

[DM04] Stefan Dziembowski and Ueli Maurer. “On Generating the Initial Key in the Bounded-Storage Model”. In: *Advances in Cryptology - EUROCRYPT 2004*. Ed. by Christian Cachin and Jan L. Camenisch. Springer Berlin Heidelberg, 2004, pp. 126–137. isbn: 1978-3-540-24676-3.

[EMMM11] David Elkouss, Jesus Martinez-Mateo, and Vicente Martin. “Information Reconciliation for Quantum Key Distribution”. In: *Quantum Information and Computation*, 11.34 (2011). arXiv: 1007.1616v2 [quant-ph].

[Erv+14] Christopher Erven et al. “An experimental implementation of oblivious transfer in the noisy storage model”. In: *Nature Communications* 5.1 (2014). doi: 10. 1038/ncomms4418.

[EGL82] Shimon Even, Oded Goldreich, and Abraham Lempel. “A randomized protocol for signing contracts”. In: *Advances in Cryptology-CRYPTO ’82*. Plenum, 1982, pp. 205–210.

[Fil] Tomer Filiba. *reedsolo* 0.3. url: https://pypi.org/project/reedsolo/.

[Gal62] Robert G. Gallager. “Low-density parity-check codes”. In: *IRE Transactions on Information Theory* 8.1 (1962), pp. 21–28. issn: 0096-1000.

[GPCZ18] Ran Gelles, Anat Paskin-Cherniavsky, and

Vassilis Zikas. *Secure Two-Party Computation over Unreliable Channels.* Cryptology ePrint Archive, Report 2018/506. https://eprint.iacr.org/2018/506. 2018.

[GV88] Oded Goldreich and Ronen Vainish. "How to solve any protocol problem - an efficiency improvement (extended abstract)". In: *Advances in Cryptology — CRYPTO '87*. Ed. by Carl Pomerance. Vol. 293. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 1988, pp. 73–86. doi: 10.1007/3-540-48184-2_6.

[Got97] Daniel Gottesman. "Stabilizer Codes and Quantum Error Correction". In: (May 28, 1997). arXiv: quant-ph/9705052v1 [quant-ph]. [GJC12] Daniel Gottesman, Thomas Jennewein, and Sarah Croke. "Longer-Baseline Telescopes Using Quantum Repeaters". In: *Physical Review Letters* 109.7 (2012). doi: 10.1103/physrevlett.109.070503.

[JNN12] J.R. Johansson, P.D. Nation, and Franco Nori. "QuTiP: An open-source Python framework for the dynamics of open quantum systems". In: *Computer Physics Communications* 183.8 (2012), pp. 1760–1772. doi: 10.1016/j.cpc.2012.02. 021.

[Kil88] Joe Kilian. "Founding cryptography on oblivious transfer". In: *Proceedings of the Twentieth Annual ACM Symposium on Theory of Computing*. STOC '88. ACM, 1988, pp. 20–31. doi: 10.1145/62212.62215.

[Kim08] H. J. Kimble. "The quantum internet". In: *Nature* 453 (2008), pp. 1023–1030. doi: 10.1038/nature07127. arXiv: 0806.4195v1.

[Kóm+14] Peter Kómár et al. "A quantum network of clocks". In: *Nature Physics* 10.8 (2014), pp. 582–587. doi: 10.1038/nphys3000.

[KWW12] Robert König, Stephanie Wehner, and Jürg Wullschleger. "Unconditional security from noisy quantum storage". In: *IEEE Transactions on Information Theory* 58.3 (2012), pp. 1962–1984. doi: 10.1109/TIT.2011.2177772. arXiv: 0906.1030v4.

[LC83] Shu Lin and Daniel J. Costello. *Error Control Coding: Fundamentals and Applications*. Prentice Hall, 1983. isbn: 013283796x.

[Lo97] Hoi-Kwong Lo. "Insecurity of quantum secure computations". In: *Physical Review* A 56.2 (1997), pp. 1154–1162. doi: 10.1103/PhysRevA.56.1154.

[LC97] Hoi-Kwong Lo and Hoi Fung Chau. "Is quantum bit commitment really possible?" In: *Physical Review Letters* 78.17 (1997), pp. 3410–3413. doi: 10.1103/PhysRevLett.78.3410.

[Mac03] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Pr., Sept. 1, 2003. isbn: 0521642981.

[May97] Dominic Mayers. "Unconditionally secure quantum bit commitment is impossible". In: *Physical Review Letters* 78.17 (1997), pp. 3414–3417. doi: 10.1103/PhysRevLett.78.3414.

[MT13] Rodney Meter and Joe Touch. "Designing quantum repeater networks". In: *IEEE Communications Magazine* 51.8 (2013), pp. 64–71. doi: 10.1109/mcom.2013. 6576340.

[Net] NetSquid. url: https://netsquid.org/.

[NC10] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Pr., Dec. 1, 2010. isbn: 1107002176.

[Pyt] Python. url: https://www.python.org/.

[QuT] QuTech. *Quantum Internet and Networked Computing*. url: https://qutech.nl/roadmap /quantum-internet/.

[Rab81] Michael O. Rabin. *How to exchange secrets by oblivious transfer.* Technical Report TR-81. Harvard University, 1981.

[RS60] I. S. Reed and G. Solomon. "Polynomial Codes Over Certain Finite Fields". In: *Journal of the Society for Industrial and Applied Mathematics* 8.2 (1960), pp. 300– 304. doi: 10.1137/0108018.

[RSA78] R. L. Rivest, A. Shamir, and L. Adleman. "A method for obtaining digital signatures and public-key cryptosystems". In: *Communications of the ACM* 21.2 (1978), pp. 120–126. doi: 10.1145/359340.359342.

[Sca+09] Valerio Scarani et al. "The Security of Practical Quantum Key Distribution". In: *Reviews of Modern Physics* 81.3 (2009), pp. 1301–1350. doi: 10 . 1103 / RevModPhys.81.1301. arXiv: 0802.4155v3.

[Sch10] Christian Schaffner. "Simple protocols for oblivious transfer and secure identification in the noisy-quantum-storage model". In:

*Physical Review A* 82.3 (2010), p. 032308. doi: 10.1103/PhysRevA.82.032308. arXiv: 1002.1495v2.

[STW09] Christian Schaffner, Barbara M. Terhal, and Stephanie Wehner. “Robust cryptography in the noisy-quantum-storage model”. In: *Quantum Information and Computation* 9.11-12 (2009), pp. 963–996. arXiv: 0807.1333v3.

[Sha48] Claude E. Shannon. “A mathematical theory of communication”. eng. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. issn: 0005-8580.

[Sho95] Peter W. Shor. “Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer”. In: *SIAM J.Sci.Statist.Comput. 26 (1997) 1484* (Aug. 30, 1995). doi: 10.1137/S0097539795293172. arXiv: quantph/9508027v2 [quant-ph].

[Sim] SimulaQron. url: http://simulaqron.org/.

[SHT18] Damian S. Steiger, Thomas Häner, and Matthias Troyer. “ProjectQ: an open source software framework for quantum computing”. In: *Quantum 2* (2018), p. 49. doi: 10.22331/q-2018-01-31-49.

[TL17] Marco Tomamichel and Anthony Leverrier. “A largely self-contained and complete security proof for quantum key distribution”. In: *Quantum 1* (2017), p. 14. Doi: 10.22331/q-2017-07-14-14.

[Twi] Twisted. url: https://twistedmatrix.com/trac/.

[Weh08] Stephanie Wehner. “Cryptography in a Quantum World”. In: (2008). arXiv: 0806.3483v1 [quant-ph].

[WEH18] Stephanie Wehner, David Elkouss, and Ronald Hanson. “Quantum internet: a vision for the road ahead”. In: *Science* 362.6412 (2018). doi: 10.1126/science.aam9288.

[WST08] Stephanie Wehner, Christian Schaffner, and Barbara M. Terhal. “Cryptography from noisy storage”. In: *Physical Review Letters* 100.22 (2008), p. 220502. doi: 10.1103/PhysRevLett.100.220502. arXiv: 0711.2895v3.

[Weh+10] Stephanie Wehner et al. “Implementation of two-party protocols in the noisy-storage model”. eng. In: *Physical Review A: Atomic, Molecular and Optical Physics* 81 (2010), urn:issn:1050–2947. issn: 1050-2947.

[WZ82] William K. Wootters and Wojciech H. Zurek. “A single quantum cannot be cloned”. In: *Nature* 299.5886 (1982), pp. 802–803. doi: 10.1038/299802a0.

[Yao82] Andrew C. Yao. “Protocols for secure computations”. In: *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982).* IEEE, 1982. doi: 10.1109/sfcs.1982.38.

[Zho+12] Zong-Quan Zhou et al. “Realization of Reliable Solid-State Quantum Memory for Photonic Polarization Qubit”. In: *Physical Review Letters* 108.19 (2012). doi: 10.1103/physrevlett.108.190505.

Sciences

# Restricted Boltzmann Machines and the Renormalization Group:
## *Learning Relevant Information in Statistical Physics*

Jesse Hoogland
*Supervisor*
Dr. P. Marcos Crichigno (UvA)
*Reader*
Prof. Dr. Max Welling (UvA)

### Abstract

Recent work has drawn attention to the links between statistical physics and machine learning (ML) and, in particular, to comparisons between the renormalization group (RG) and deep neural networks, respectively. These have inspired renewed interest in the information-theoretic framework underpinning these fields, prompting a better understanding of what RG is. In this capstone, we introduce and expand upon these connections from the ground up. Starting with the basics of ML and RG, we work our way to an algorithm implemented on neural networks that learns optimal, model-independent RG procedures: the real-space mutual information (RSMI) algorithm. In doing so, we review the current state of the literature, clarifying misconceptions in earlier works. With the RSMI algorithm, we review a novel calculation of the Ising model critical exponent and generalize this approach to arbitrary lattice systems. We release an open-source library, *rgpy*, for implementing these novel procedures, and close with a discussion of the wide-ranging implications.

Keywords and phrases: *machine learning, restricted Boltzmann machines, the renormalization group, information theory, mutual information*

## Acknowledgements

I would like to thank Marcos Crichigno for supervising this capstone. I do not think either of us knew what we were getting into when I first approached him, and I am stunned at how much I have been able to learn during this time. Marcos provided guidance, much-needed prodding, and a watchful eye, keeping the focus on the big picture. The topics were new to both of us, ML for him and RG for me, and getting to integrate these different perspectives was a wonderful experience. I would also like to thank my dad, Jiri Hoogland, for helpful feedback on drafts and, most of all, for his expertise on on all things Monte Carlo. After being stuck with implementation details of the RSMI algorithm for two weeks, his words finally let the penny drop. As a last note, I would like to thank my friends and (other) family for putting up with me during the last weeks (when I was a little less than sociable), in particular, Nick van der Woude, for mutually infectious enthusiasm.

## I. Introduction

In the age of *big data*, machine learning (ML), a subset of artificial intelligence (AI), has become more than *just* another set of data analysis tools [1]. First, ML's connections with theoretical physics are multivarious and deeply conceptual; the very success of ML may, in part, result from physical principles including symmetry, locality, and hierarchy [2]. Furthermore, ML and theoretical physics share a powerful conceptual framework in information theory [3]. Beyond data analysis, the intersection of ML and physics contains a unique set of ideas that researchers in both fields can leverage to solve tough problems.

In particular, recent work has drawn attention to the similarities between ML and a class of techniques from statistical physics known as the renormalization group (RG) [4, 5, 6, 2]. Developed in the last century, RG has been crucial in making sense of critical behavior – those phenomena characterizing phase transitions. In 2014, Mehta and Schwab published a seminal paper describing an *exact* equivalence between a technique from RG and a type of neural network (NN) from ML [4]. This, however, encountered criticism, and it works only under a narrow set of circumstances. The similarities between ML and RG, then, are still largely qualitative, and this remains an active area of research. In addition, the research landscape maintains lingering misconceptions about the details of the intersection [6, 2, 7, 8]. This warrants further investigation, and in order to facilitate and encourage such research, our first contribution is to provide a clarifying overview of the competing views; in doing so we resolve a number of inaccuracies.

In 2018, Koch-Janusz and Ringel derived an algorithm which uses neural networks to learn RG transformations on lattice systems: the *real-space mutual information* (RSMI) algorithm [5]. Notably, this method is *unsupervised*, which is particularly relevant for research into poorly understood physical systems; information-theoretic approaches like this algorithm may guide researchers towards the locations of critical points and even calculations of critical exponents. Furthermore, Koch-Janusz and Ringel's derivation is *optimal* in the rigorous sense we define in Chapter 5 [5, 9]. This is exciting because many well-established practices in RG lack precise justification. More exact formulations like these may inspire more effective implementations, not to mention a better understanding of why these ML and RG techniques work.

In our investigation, we will develop a set of tools for tackling critical phenomena. First, we consider some of the standard techniques of statistical physics [2], building towards an ML-derived implementation of RG [5]. Second, we introduce elements of ML, emphasizing their utility in a variety of statistical physics contexts [3]. We anchor this investigation around the Ising model, a description of ferromagnetism and one of the most important models in statistical physics. To compare these various techniques, we evaluate their ability in predicting the Ising model's correlation length critical exponent, $\nu$.

To accomplish this, we have built and shared an implementation of the RSMI algorithm [10] in the open-source Python library *rgpy* [1] [6]. Hereby, we provide a calculation for $\nu$ [6]. Then, we describe a generalization of this algorithm to *arbitrary* lattice systems, giving rise to a family of RSMI-inspired approaches.

It is our aim to enable and inspire researchers to build further on our results. We accomplish this by reviewing the current state of research, sharing an implementation of the RSMI algorithm, and describing avenues of future research. Although we focus on the perspective of statistical physicists, this capstone is accessible for both ML and physics researchers, even at the undergraduate level.

In Chapter 2, we begin by introducing techniques native to statistical physics. We describe mean-field theory and its failures, which brings us to the renormalization group. In Chapter 3, we discuss two examples of ML in physical investigations. First, we use neural networks to classify Ising model phases. Then, we use the same neural networks to generate new samples of Ising models. These examples serve to introduce the basics of ML, assuming no prior knowledge (except mathematical maturity), and the same is true for the portion on statistical physics. In Chapter 4, we explore the similarities between ML and RG, and by being more explicit in how we define "relevant" information, we achieve a more precise comparison between the two. In Chapter 5, we explain and justify the RSMI algorithm, following the formulation of Koch-Janusz and Ringel. In Chapter 6, we provide our own results: a recalculation of $\nu$ and a generalization of this technique, paving the way for a new class of RG techniques. In Chapter 7, we close with a discussion, reflecting on our comparisons of ML and statistical physics and emphasizing the wide-ranging impacts of these ideas.

### Notation

To refer to single microscopic elements (e.g. spins in the Ising model or pixels in an image), we use lowercase letters with a lower index ($x_i$, $y_j$, etc.). To refer to collections of microscopic elements, microstates or images, we use boldface, lowercase letters ($\boldsymbol{x} \coloneqq \{x_i\}$, $\boldsymbol{y} \coloneqq \{y_j\}$, etc.).

To refer to collections of microstates, we use uppercase, cursive letters, $\mathcal{X} \coloneqq \{\boldsymbol{x}\}$. We will be interested in performing sums and averages over these sets. Rather than introduce an index to keep track of each term, we do so implicitly in the sums. For example, given some function $A(\boldsymbol{x})$, the following are equivalent: $\sum_n A(\boldsymbol{x}^{(n)}) \equiv \sum_{\boldsymbol{x}\in\mathcal{X}} A(\boldsymbol{x}) \equiv \sum_{\boldsymbol{x}} A(\boldsymbol{x})$. Most often, we use the last notation.

If we partition our microstates into subsets (as with block renormalization), we also use boldface, lowercase letters ($\boldsymbol{v}$, $\boldsymbol{h}$, etc.). To distinguish partitions, we may use an upper index: $\boldsymbol{v}^{(n)}$.

For partial derivatives, we typically use the shorthand $\partial_t \coloneqq \frac{\partial}{\partial t}$.

For Ising models, we will consider systems with binary units $\in \{-1, 1\}$, following standard convention. For RBMs, we use the standard notation of binary units $\in \{0, 1\}$. When using RBMs on Ising data, then, we map $-1 \to 0$.

## II. Foundations of Statistical Physics

Statistical physics emerged in the second half of the nineteenth century as an answer to unresolved questions in thermodynamics, the study of heat and work. Was heat continuous and wavelike, or might it be something else, discrete and atomic? Founding figures in the field, such as Rudolf Clausius, Ludwig Boltzmann, and James Clerk Maxwell, answered the latter. Introducing the kinetic theory of gases, these scientists posited gases as large collections of tiny molecules and heat flow as the net effect of unbalanced molecular collisions. By translating these mi-

[1] A programming library contains a set of functionalities easily accessible to other computer scientists.

croscopic descriptions to experimentally-verifiable, macroscopic predictions, these physicists were able to defend this theory of gases. The techniques pioneered in performing this translation would give rise to the field of statistical physics [11].

To complicate matters, these scientists lacked equipment that could resolve the proposed microscopic length scales, and a square cubic centimeter of gas can contain upwards of a million million million molecules [11]. The key insight in statistical physics is to focus on the properties of the collection rather than on the individual components – on averages and distributions rather than microscopic details. The translation between microscopic and macroscopic is the essence of statistical physics, and it is to this task we dedicate our efforts.

Our investigation begins by defining a *microstate*, $\boldsymbol{s}$: a full description of the microscopic *degrees of freedom* of our system, $\boldsymbol{s} \coloneqq \{s_i\}$, where $s_i$ is the $i$-th DOF, some fundamental way in which the system can vary.[2] Our aim, as statistical physicists, is to predict the outcomes of macroscopic measurements. A key assumption of statistical mechanics is that we can express measurement outcomes as averages over all microstates, $\mathcal{S}$ (see Appendix A.1). If we are interested in measuring energy, $E$, our *expectation* $\langle E \rangle$, will be:

$$\langle E \rangle \coloneqq \sum_{\boldsymbol{s} \in \mathcal{S}} P(\boldsymbol{s}) E(\boldsymbol{s}) \tag{2.1}$$

Making macroscopic predictions boils down to evaluating probabilities of microstates and the sums thereof. In the following section, we will derive the probability distribution $P(\boldsymbol{s})$ and encounter the first of the fundamental challenges in statistical physics, following the treatment of Domb [11] and Cardy [12].

## Probabilities and Partition Functions

Let us consider an example: we begin with some physical system, $\mathcal{S}$. It could be metal, gas, or another material. As we previewed, the microscopic details are irrelevant to the macroscopic picture. To measure macroscopic properties of $\mathcal{S}$, we need to know its distribution over microstates $\boldsymbol{s}$, $P(\boldsymbol{s})$. The trick is to introduce a *reservoir*, $\mathcal{R}$, that surrounds $\mathcal{S}$. Just like $\mathcal{S}$, the reservoir could be anything: a gas, liquid, etc., where we impose two conditions: $\mathcal{S}$ and $\mathcal{R}$ exchange only energy, and the combined system, $\mathcal{X} = \mathcal{S} \cup \mathcal{R}$ (the *universe*), is isolated. Then, the total energy, $E$, is conserved. If we denote the energy of microstates, $\boldsymbol{s}$ and $\boldsymbol{r} = \{r_j\}$ as $E(\boldsymbol{s})$ and $E(\boldsymbol{r})$, respectively, energy conservation requires $E(\boldsymbol{s}) + E(\boldsymbol{r}) = E$.[3] Furthermore, we treat $\mathcal{S}$ as a much smaller fraction of $\mathcal{X}$ than $\mathcal{R}$, though $\mathcal{S}$ is still macroscopic ($1 \ll |\boldsymbol{s}| \ll |\boldsymbol{r}|$, where $|\boldsymbol{x}|$ denotes the number of degrees of freedom of $\boldsymbol{x}$).

The fundamental assumption of statistical mechanics states that, for an isolated system, each microstate is equally probable. Then, our probability distribution is $P(\boldsymbol{x}) = 1/\Omega(\boldsymbol{x})$, where $\Omega(\boldsymbol{x})$ is the number of possible microstates $\boldsymbol{x}$. Probabilities over subsets of such a system may be more complicated. Consider that the probability of $\boldsymbol{s}$ is proportional to the number of ways we can rearrange the reservoir, keeping the energy constant. If we let $\Omega_{\boldsymbol{r}}(\boldsymbol{s})$ denote the number of microstates, $\boldsymbol{r}$, with this energy $E - E(\boldsymbol{s})$:

$$P(\boldsymbol{s}) = c\Omega_{\boldsymbol{r}}(\boldsymbol{s}), \tag{2.2}$$

where $c$ is the constant of proportionality. By the fundamental assumption of statistical mechanics, the above holds for any choice in microstate, $\boldsymbol{s}'$:

$$P(\boldsymbol{s}') = c\Omega_{\boldsymbol{r}}(\boldsymbol{s}'). \tag{2.3}$$

Although we do not have enough information to evaluate absolute probabilities, we can now compare the relative likelihood of different microstates.

$$\frac{P(\boldsymbol{s})}{P(\boldsymbol{s}')} = \frac{\Omega_{\boldsymbol{r}}(\boldsymbol{s})}{\Omega_{\boldsymbol{r}}(\boldsymbol{s}')}. \tag{2.4}$$

To transform the above into a more manageable form, we introduce the Boltzmann Entropy:

$$S(\boldsymbol{S}) \coloneqq k \log \Omega(\boldsymbol{S}), \tag{2.5}$$

[2] For example, the concept of *spin*, which we will encounter in the next section.

[3] Let us consider how $\boldsymbol{s}$ and $\boldsymbol{r}$ might exchange energy physically. If $\boldsymbol{s}$ is a metal and $\boldsymbol{r}$ a gas, then the two will transfer energy whenever gas particles collide against the metal, exchanging energy stored in the metal's vibrations with the kinetic energy of moving gas particles.

where $k$ is Boltzmann's constant, a scaling factor from the microscopic to macroscopic. Another crucial definition is that of *macrostate*: a collection of microstates, $\boldsymbol{S} \subset \mathcal{S}$, indistinguishable to the experimental observer. Whereas this observer can measure differences between different macrostates, the differences within a given macrostate are nondifferentiable. For the statistical physicist, translating microscopic predictions to macroscopic predictions amounts to making statements about how probable different macrostates are as a function of how probable the underlying microstates are. To facilitate this, we introduce $\Omega(\boldsymbol{S})$, the number of microstates corresponding to a macrostate, $\Omega(\boldsymbol{S}) = |\boldsymbol{S}|$. We interpret $S(\boldsymbol{S})$ as a measure of uncertainty: higher entropy gives a lower probability of correctly guessing the true microstate the system occupies, which can complicate the translation between microstates and macrostates.

Returning to the task at hand, we can plug our equation for the entropy of the reservoir into equation (2.4). Subsequently, we get:

$$\frac{P(\boldsymbol{s})}{P(\boldsymbol{s}')} = e^{\frac{1}{k}(S_{\boldsymbol{r}}(\boldsymbol{s}) - S_{\boldsymbol{r}}(\boldsymbol{s}'))}. \tag{2.6}$$

By our assumption that the reservoir is much larger than $\mathcal{S}$, we can approximate the above using the second law of thermodynamics, $T\Delta S \approx \Delta E$ (constant volume and number of particles), to get: [4]

$$\frac{P(\boldsymbol{s})}{P(\boldsymbol{s}')} = e^{-\beta(E_{\boldsymbol{s}}(\boldsymbol{s}) - E_{\boldsymbol{s}}(\boldsymbol{s}'))}, \tag{2.7}$$

where $\beta\beta := 1/(kT)$, the thermodynamic beta. This requires that $\mathcal{S}$ and $\mathcal{R}$ are in thermal equilibrium, i.e. their temperatures are the same (for a more thorough justification of these steps, see Appendix A.2). By the fact that $\boldsymbol{s}$ and $\boldsymbol{s}'$ are independent, we can separate (2.7) to get that:

$$P(\boldsymbol{s}) \propto e^{-\beta E(\boldsymbol{s})}. \tag{2.8}$$

By normalizing (solving $\sum_{\boldsymbol{s}} P(\boldsymbol{s}) = 1$), we get an equation for the probability distribution, the so-called Boltzmann distribution:

$$\boxed{P(\boldsymbol{s}) := \frac{1}{Z} e^{-\beta E(\boldsymbol{s})} \qquad Z := \sum e^{-\boldsymbol{E}(\boldsymbol{s})}} \tag{2.9}$$

where $Z$ is the normalizing factor, *the partition function*. It holds for any system $\boldsymbol{s}$ so long as $\boldsymbol{s}$ exchanges only energy with its surroundings. Together, these conditions—fixed temperature, number of particles, and volume—form the *canonical ensemble*. Within the canonical esemble we have found a relation between the energy of a microstate and the probability that a system occupies that microstate at any given time. If we specify an energy function we can calculate probabilities and, from those probabilities, our desired measurement outcomes. In Chapter 3, we will see this equation again, describing the evolution of artificial neurons in neural networks. With the appropriate choice in energy function, equation (2.9) helps characterize certain neural networks, and much of our subsequent analysis will translate readily.

### Ising Model

An example of a system with a suitable energy function for the Boltzmann distribution is the Ising model depicted in Figure 2.1. In this model, we require that the microscopic degrees of freedom $s_i$ are binary-valued ($s_i \in \{1, -1\}$), and we call these degrees of freedom *spins*. The Ising model was conceived as a minimal model for *ferromagnetism*, the phenomenon by which metals form permanent magnets. In nature, spin is an intrinsic property of particles that induces and interacts with magnetic fields. Though it is a quintessentially quantum effect, we can approximate spin classically as orienting either "up" or "down" ($+1$ and $-1$ respectively).

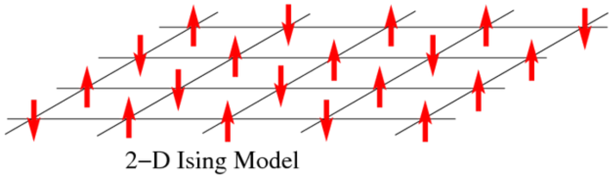


*Figure 2.1 – The Ising model is a minimal model for ferromagnetism. Image source [13].*

We write the following *Hamiltonian* (energy function) for the Ising model:

[4] For other statistical ensembles, we might include other terms, like the number of particles.

$$E(\boldsymbol{s}) := -B\sum_i s_i - J\sum_{\langle i,j\rangle} s_i s_j, \tag{2.10}$$

where, in the ferromagnet, $B$ is the external magnetic field, $J$ is the interaction energy between neighboring pairs of spins, and $\sum_{\langle i,j\rangle}$ denotes a sum over adjacent sites. We see that the system is in a lower energy state when spins $s_i$ align* with $B$ and their neighbors $J\sum_{j\to i} s_j$, where $j \to i$ denotes the neighbors of $i$.

### Intractable sums

For the vast majority of conceivable Hamiltonians, equation (2.9) is intractable. This stems from $Z$, the *partition function*. For an Ising magnet with $N$ spins, $Z$ will contain $2^N$ terms. Beyond around $N = 300$, this exceeds the number of atoms in the universe [14]. In the standard thermodynamic limit that $N$ goes to infinity, this diverges. Even in everyday (finite) life, $N$ is on the order of Avogadro's number so is already incredibly large. How are we to proceed? To complicate matters further, equation (2.1) requires another sum of $2^n$ terms. It turns out that with regard to this last quandary, $Z$ will be our saving grace. $Z$, being more than *merely* a normalizing factor, contains all the relevant information about our specific system. From $Z$, we can determine any desired macroscopic parameters of interest by taking derivatives. For example, if we define the free energy, $F := -\beta \ln Z$, then for the Ising model our expectations for the magnetization will be $\langle M\rangle = \partial_B F$ (see Appendix A.3), where $M(\boldsymbol{s}) := \sum_i s_i$, the net orientation of all spins. In fact, with clever tricks, we can actually solve (i.e. add up) the infinite sums for the Ising model Hamiltonian in one or two dimensions to derive *exact* solutions for expectation values. However, this is the exception, not the rule.

Most models of interest, we cannot solve exactly. Then, the best we can do is make approximate solutions. One class of possibilities is Markov Chain Monte Carlo (MCMC) techniques. In using these techniques, rather than evaluating our sums and averages over all microstates, $\mathcal{S}$, we evaluate these over a representative, *finite* set of samples, $\mathcal{S}_{data}$ (see Figure 2.2). Furthermore, rather than evaluating our average over a lattice infinite in extent, we evaluate our average over a finite lattice. The results will, in general, differ from the true infinite sums we would like to calculate. However, as we increase the size of the lattice and the number of samples, our results can get arbitrarily close. The key trick in MCMC techniques is that relative probabilities, $P(\boldsymbol{s}')/P(\boldsymbol{s})$, are much easier to evaluate than absolute probabilities, $P(\boldsymbol{s})$, since the partition functions cancel. Monte Carlo techniques proceed according to some variation of:

1. Begin with a randomly-chosen starting state, $\boldsymbol{s}$.

2. Consider a small variation to the state, $\boldsymbol{s}'$ (for example, by flipping spin $s_i$).

3. Decide whether to accept this variation according to $P(\boldsymbol{s}')/P(\boldsymbol{s})$.

4. Repeat steps 2 and 3 until $\boldsymbol{s}'$ converges the *equilibrium distribution*, $P(\boldsymbol{s}')$.

Seeing as these methods are quite computationally intensive, earlier generations of physicists developed other sets of techniques. The first we will consider is mean-field theory (MFT).

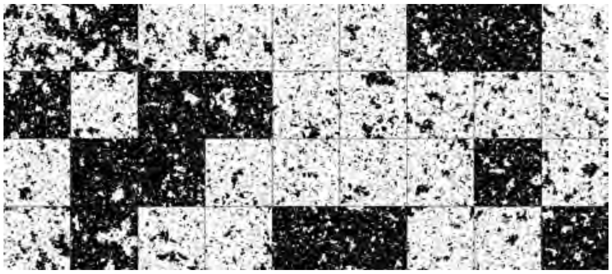


*Figure 2.2 – Samples of the 2D Ising model near the critical temperature generated with the Swendsen-Wang Algorithm, implemented rgpy.*

## Mean-Field Theory

To start, let us consider a simpler problem. If we look at a single spin, $s_i$, having specified all other spins, can we calculate its partition function? First, we can rewrite equation (2.10) more instructively:

$$E(\boldsymbol{s}) = -\sum_i E_i(s_i), \qquad E_i(s_i) := -\left(\sum_{j\to i} J s_j + B\right) s_i = -( \tag{2.11}$$

where $n$ is the number of neighbors of $s_i$ and $\langle s_{j\to i}\rangle \coloneqq \frac{1}{n}\sum_{j\to i} s_j$ is the average spin of $s_i$'s neighbors. Then, using Boltzmann's equation, the probability over $s_i$ is:

$$P(s_i) = \frac{e^{-\beta(nJ\langle s_{j\to i}\rangle + B)s_i}}{e^{-\beta(nJ\langle s_{j\to i}\rangle + B)} + e^{\beta(nJ\langle s_{j\to i}\rangle + B)}}, \tag{2.12}$$

and

$$\langle s_i\rangle = P(s_i = 1) - P(s_i = -1) = \frac{2\cosh\beta(nJ\langle s_{j\to i}\rangle + B)}{2\sinh\beta(nJ\langle s_{j\to i}\rangle + B)} = \tanh\beta(nJ\langle s_{j\to i}\rangle + B). \tag{2.13}$$

So we see that the orientation of a given spin depends on the average orientation of its neighbors, as one might expect.

If we were given $\langle s_{j\to i}\rangle$, evaluating this equation is trivial. Just as it is easy to calculate relative probabilities, it is easy to calculate conditional probabilities like $P(s_i|\{s_{j\to i}\})$. Here, too, the partition functions cancel: $P(s_i|\{s_{j\to i}\}) = P(s_i, \{s_{j\to i}\})/P(\{s_{j\to i}\})$.[5]

For now though, these observations are not of much help since we do not know $\langle s_{j\to i}\rangle$. This brings us to the mean-field approximation. Since we could have chosen any spin $s_i$ as our starting point (including its neighbors), we expect, $\langle s_i\rangle \approx \langle s_{j\to i}\rangle$, known as the principle of mediocrity. In the mean-field approximation we assume, more stringently, that $\langle s\rangle \coloneqq \langle s_i\rangle = \langle s_{j\to i}\rangle$, so:

$$\langle s\rangle = \tanh(\beta nJ\langle s\rangle + B). \tag{2.14}$$

If we find a solution using $\langle M(\boldsymbol{s})\rangle = \langle\sum_i s_i\rangle = \sum_i\langle s_i\rangle$, we would have our measurement outcome:

$$\langle M(\boldsymbol{s})\rangle = \sum_i \langle s_i\rangle = N\langle s\rangle, \tag{2.15}$$

and we see that $\langle s\rangle$ is nothing more than magnetization per site $m = \langle M\rangle/N$, where $N$, as before, is the number of spins.

It turns out that we cannot solve equation (2.14) analytically (it is a transcendental equation), so we have to resort to numerical techniques. For intuition though, we can get far with a graphical approach (see Figure 2.3).

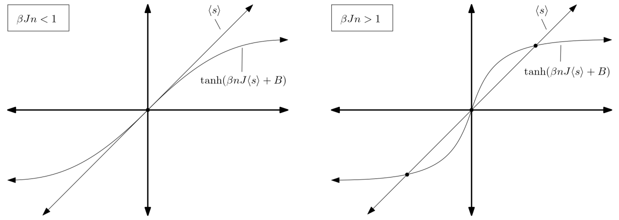


*Figure 2.3 – Mean-Field Theory predictions for the spontaneous magnetization $M|_{B=0}$.*

Restricting to the case that $B = 0$, let us distinguish two cases:

1. $\beta Jn < 1$. There is only one solution: $m = 0$.
2. $\beta Jn > 1$. Suddenly, there are two additional solutions. Something interesting seems to happen at the point $\beta Jn = 1$; we call this a *critical point*, and it occurs at the *critical temperature*, $T_c = Jn/k$.

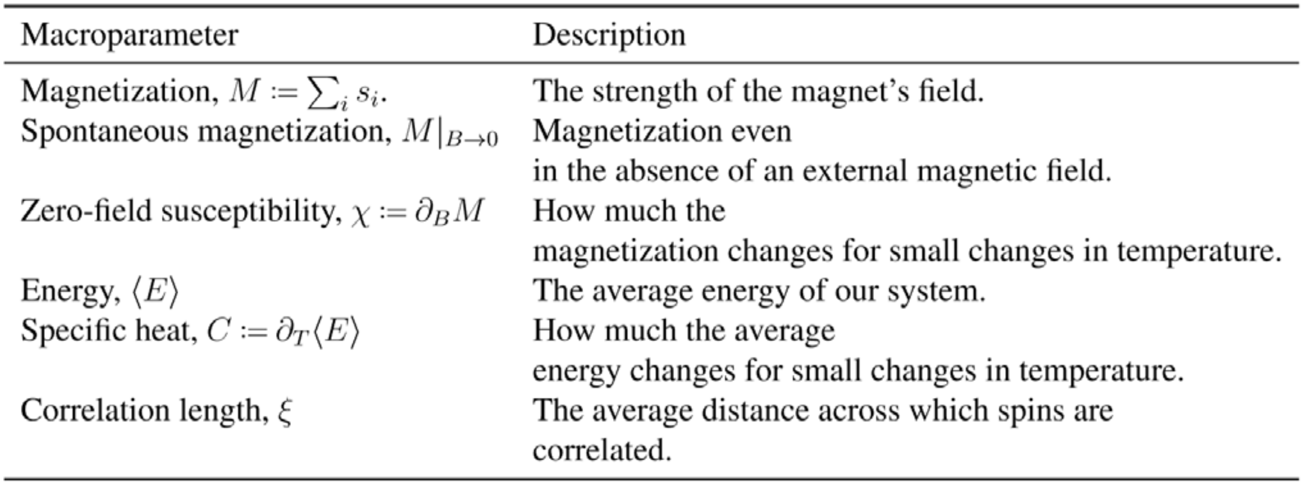
Taylor-expanding the right side of equation (2.14), we get that in the vicinity of the critical point:

$$\begin{cases} = 0 & T > T_c \sim \pm(3|t|)^{-1/2} \\ T < T_c, \end{cases} \tag{2.16}$$

where $t$ is the *reduced temperature*, $t \coloneqq (T - T_c)/T_c$. In fact, these equations contain more information about our system than the magnetization alone. We can use mean-field theory to derive other parameters like the magnetic susceptibility and specific heat (see Table 2.2, some are particular to Ising-like models, others are more general). These are the quantities that we can measure in a laboratory, so by calculating these figures we can confirm (or deny) that the Ising model is a suitable model of ferromagnetism. In this light, we see why this translation step is so crucial: it provides the means of testing our theories. Until now, our discussion assumed an arbitrary number of neighbors, $n$, so we would expect our equations for these parameters to hold for any number of dimensions. It would seem like we have accomplished our goal for this chapter a full section in advance.

[5] This avoiding of joint probabilities with clever choices in conditional probabilities will be the basis for efficiently training RBMs in the next section.

Table 2.1: Macroscopic Parameters of the Ising model



| Macroparameter | Description |
|---|---|
| Magnetization, $M := \sum_i s_i$. | The strength of the magnet's field. |
| Spontaneous magnetization, $M\|_{B\to 0}$ | Magnetization even in the absence of an external magnetic field. |
| Zero-field susceptibility, $\chi := \partial_B M$ | How much the magnetization changes for small changes in temperature. |
| Energy, $\langle E\rangle$ | The average energy of our system. |
| Specific heat, $C := \partial_T \langle E\rangle$ | How much the average energy changes for small changes in temperature. |
| Correlation length, $\xi$ | The average distance across which spins are correlated. |

Unfortunately however, in less than four dimensions, mean-field theory provides incorrect predictions: equation (2.16) is wrong. Intuitively, systems with less than four dimensions have too little order for MFT to hold. Introducing more dimensions means there are more paths between any two spins and more correlation between them. Past the *critical dimension* of four, there is enough order for our MFT approximation to hold. However, for the Ising model in 2 and 3 dimensions (those cases most interesting to our daily lives) we have to resort to a different approximation approach.

## Critical Phenomena and the Renormalization Group

Although the quantitative predictions of the mean-field theory are incorrect, its qualitative predictions are instructive, particularly those that predict the existence of a critical point. From equation (2.16), we expect a phase transition between a paramagnetic, disordered phase and a ferromagnetic, ordered phase (in which the system spontaneously magnetizes). This bears out experimentally, although not at the predicted temperature, see Figure 2.4.

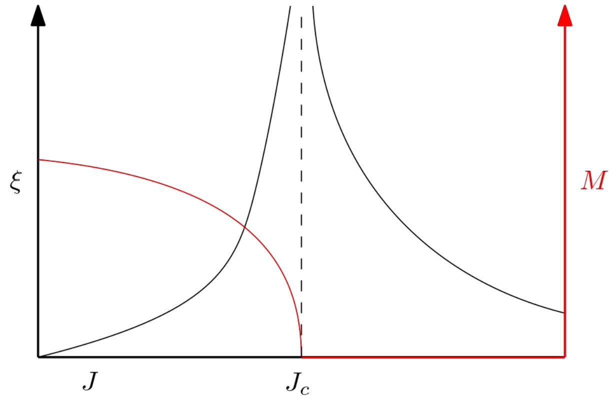


*Figure 2.4 – The qualitative behavior of the correlation length, $\xi$, and magnetization, $M$, around the critical point.*

### Correlation Length

An important quantity in Table 2.2 is the correlation length, $\xi$. This is the average distance across which spins in the system under investigation tend to fluctuate together. Spins farther apart than $\xi$ are effectively independent of one another, so severing such a connection has no appreciable effect on the macroscopic properties: thus we can think of $\xi$ as a measure of how macroscopic our system is [12]. Mean-field theory predicts that near the critical point the correlation length scales as:

$$\boxed{\xi \sim |t|^{-1/2}} \tag{2.17}$$

Although the exponent does not line up with experimental results as the true value is $1$, mean-field theory correctly predicts that $\xi$ diverges at the critical point (see Figure 2.4). In fact, this is the defining characteristic of critical points. When the correlation length diverges the entire system becomes correlated. Any perturbation, no matter how infinitesimal, will have macroscopic ramifications. For the statistical physicist, critical points are excellent places to test theories as they allow closer access to the microscopic realm.

### Critical Exponents

Another valuable prediction of mean-field theory is that of *critical exponents*. We see from equation (2.17) and equation (2.16) that near the critical point the correlation length and the magnetization obey simple power-laws. These are examples of a more general trend: near critical points, macroparameters will follow power-law scaling formulas. We call the exponents that define these relations *critical exponents* (see Table 2.3).

We shift our goal to the (correct) calculation of these critical exponents. To this end, we turn to the renormalization group, a set of ideas for tackling precisely these critical phenomena.

### The Renormalization Groups

Instead of trying to compute $Z$ head-on, let us consider a different angle. We will try to re-express $Z$ with a simpler set of parameters while preserving the physical, long-distance information. Repeating these transformations, we will discard the irrelevant, microscopic fluctuations, keeping only the macroscopic information. Formally, an RG transformation will look something like:

$$\sum_{\boldsymbol{s}'} e^{-H'(\boldsymbol{s}')} = \sum_{\boldsymbol{s}} e^{-H(\boldsymbol{s})}, \tag{2.18}$$

constraining, for example, $|\boldsymbol{s}'| < |\boldsymbol{s}|$. We consider $H'$ and $H$ parameterized with sets of couplings $\{K'\}$ and $\{K\}$, for example,

$$H(\boldsymbol{s}) = -\sum_i K_i^{(1)} s_i - \sum_{\langle i,j \rangle} K_{ij}^{(2)} s_i s_j - \sum_{\langle\langle i,j \rangle\rangle} K_{ij}^{(3)} s_i s_j \ldots, \tag{2.19}$$

where $\langle\langle i,j \rangle\rangle$ denotes next-nearest neighbors, and the continued sum will, in general, contain all possible interactions. This equation describes all possible physical models and lattice. To recover the Ising Hamiltonian, we would choose $K_i^{(1)} = \beta B$, $K_{ij}^{(2)} = \beta J$, and all $K^(\nu)$=0$for\nu$>2.

Alone, equation (2.18) is not enough of a requirement. A "good" RG transformation satisfies a special set of criteria: it should preserve long-distance information while discarding short-distance information. Arbitrary transformations satisfying equation (2.18) need not extract the information we deem *relevant*. Formally, we are interested in extracting the *relevant operators*, those describing macroscopic properties, and suppressing the *irrelevant operators*, those describing microscopic properties. For example, we might accomplish this transformation by summing over even spins (known as *decimation*):

$$e^{-H'(\boldsymbol{s}')} = \sum_{s_2, s_4, \ldots, s_N} e^{-H(\boldsymbol{s})}, \tag{2.20}$$

where now $\boldsymbol{s}'$ ranges over the odd spins. In other words, we *integrate out* or *marginalize over* the short-distance degrees of freedom.

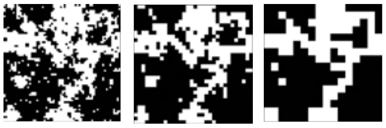


*Figure 2.5 – Three steps of majority-rule block-spin renormalization, preceding left to right (block size $b = 2$)*

Consider, first, a descriptive example: (majority-rule) block-spin renormalization, a set of RG techniques intended for lattice systems. For a given microstate, block renormalization proceeds as follows, (see Figure 2.5):

1. We partition the configuration into non-overlapping blocks. For each block, we determine which spin is in the majority and we assign that value to a new, single spin. These define a new coarse-grained system.

2. We rescale the coarse-grained configuration so that each block takes the size of an original spin.

Formally, we can write the block transformation rule as:

$$e^{-H'(\boldsymbol{s}')} := \sum_{\boldsymbol{s}} \prod_{\text{blocks}} \pi(s'; s_i) e^{-H(\boldsymbol{s})}, \tag{2.21}$$

where $\pi$ is the projection operator implementing the majority rule

$$\pi(s', s_1, \ldots, s_9) := \begin{cases} 1, & \text{if} s' = \operatorname{sgn} \sum_i s_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.22}$$

Though we can easily perform this procedure for any individual configuration, equation (2.21) requires us to do this for all microstates, and this remains intractable. Ultimately, for most systems, RG will use variational schemes [12]. It is, however, the qualitative insights RG offers, in spite of these approximations, which merit our immediate attention.

Consider what happens when we apply block renormalization to Ising configurations at different temperatures as in Figure 2.6. We see three trajectories of block renormalization for three different initial temperatures: below the critical point, at the critical point, and above it. Our first observation is that these transformations induce flows away from the critical temperature. There are three fixed points ($T = 0$, $T = T_c$, and $T = \infty$) for the Ising model, and, indeed, macroscopically these correspond to the three phases (ordered, critical, and disordered).

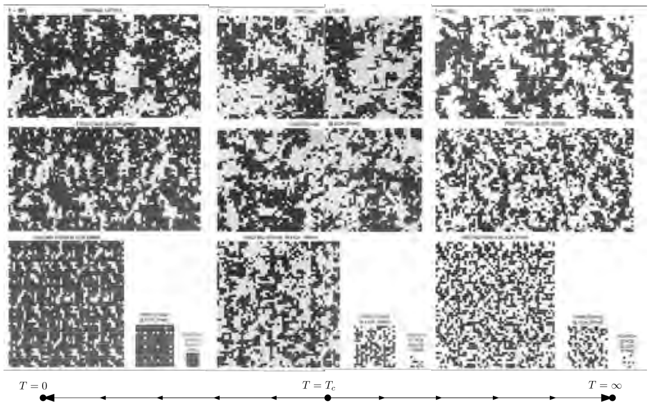


*Figure 2.6 – Renormalization induces changes in the effective temperature towards fixed points. Images from Wilson [17].*

Let us be more exact and determine how this behavior might arise. We will write the RG transformation rule as a function $\mathcal{R}$ of couplings: $\{K'\} = \mathcal{R}(\{K\})$. With two general assumptions we can build a descriptive theory of RG. First, we assume that there exists a fixed temperature point (or multiple), as would be the case with any other couplings. These are the specifications of $\{K^*\}$ that are stable under our transformation rule, namely, $\{K^*\} = \mathcal{R}(\{K^*\})$. Second, we assume we can differentiate this transformation near the fixed point, so we can linearize[6]:

$$K'_a \sim K^*_a + \sum_b \boldsymbol{\mathcal{J}}_{ab}(K_b - K^*_b), \tag{2.23}$$

where $\mathcal{J} = \frac{\partial K'_a}{\partial K_b}|_{K=K^*}$. This is the Jacobian, a generalization of the derivative to vector-valued functions of multiple variables. We thus denote $\boldsymbol{\mathcal{J}}$'s left eigenvectors $e^i$, corresponding to eigenvalues, $\lambda^i$ so that

$$\sum_a e^i_a \boldsymbol{\mathcal{J}}_{ab} = \lambda^i e^i_b.^7 \tag{2.24}$$

Next, we define *scaling variables*, $u_i := \sum_a e^i_a (K_a - K^*_a)$, which are combinations of deviations $K_a - K^*_a$ that transform multiplicatively near the fixed point:

[6] Taylor-expansions are one of the first steps in any physicists' toolkit. Here, we perform the equivalent of a Taylor expansion for a transformation of multiple variables.

[7] There is no reason to assume $\mathcal{J}$ to be symmetric or even to have real eigenvalues though we will restrict ourselves to considering real-eigenvalued Jacobians.

$$u_i' = \sum_a e_a^i (K_a' - K_a^*) = \sum_{a,b} e_a^i \mathcal{J}_{ab}(K_b - K_b^*) \quad = \sum_b \lambda^i e_b^i (K_b - K_b^*) = \lambda^i u_i$$

(2.25 & 2.26, respectively)

For later convenience, we introduce $\lambda_i \equiv b^{y_i}$, where $b$ is the rescaling size (the width of the blocks in block-spin RG). These are the *renormalization group eigenvalues* and are distinguished in three cases:

- $y_i > 0$: $u_i$ is *relevant*. Repeated RG iterations drive $u_i$ away from its fixed point value.
- $y_i < 0$: $u_i$ is *irrelevant*. Repeated RG iterations drive $u_i$ towards $0$.
- $y_i = 0$: $u_i$ is *marginal*. The linearized equations are not enough to tell us about $u_i$'s behavior.

From this, we see that our ability to distinguish between microscopic and macroscopic is a consequence of simple dimensional analysis: there are finitely many relevant eigenvalues. Those are the ones you see when you zoom out far enough. The irrelevant eigenvalues span a *critical surface* of points which, under RG transformations, are attracted towards the critical point. Macroscopically, points on this hypersurface are indistinguishable, and their behavior is fully characterized by the critical point alone. This brings us to the remarkable principle known as *universality*.

### Universality

In the last century, experimentalists faced a puzzling situation. In their efforts to measure more precise critical exponents for all manner of systems, they discovered that their experimental set-ups did not matter: all ferromagnets for a given number of dimensions possessed the same critical exponents and so, too, for all superfluids [11]. The exponents are *universal*. The Ising model, then, is not only a minimal model for ferromagnetism but can describe fluids, neural networks, metal alloys, and more.

Another key result of the RG formalism is that we can express all of our critical exponents in terms of the relevant RG eigenvalues. First, we use RG to derive a scaling rule for the free energy (see Appendix B.2). Then, from its derivatives, we can determine the critical exponents which even allows us to relate the exponents to one another in *scaling relations*. These relations had been postulated before the advent of RG but many only as inequalities. RG provided a rigorous means to link these different exponents.

For example, we show a derivation in (see Appendix B.3) for the correlation length critical exponent:

$$\boxed{\nu = \frac{1}{y_t}} \tag{2.27}$$

where $y_t$ is the thermal RG eigenvalue which is related, as its name suggests, to the temperature of the system. From the exact solution (of the Ising model in 2D), we get that $y_t = 1$. Then, we derive for the correlation length critical exponent:

$$\boxed{\nu = 1} \tag{2.28}$$

In general, since most of the sums we encounter are not exactly solvable, we do not have access to solutions like these. Therefore, we consider three approximate schemes.

### $4 - \epsilon$ Expansion.

We can rephrase the thermodynamics limit (in which we let the number of spins go to infinity) as the limit in which we hold the total size of the system fixed while letting the distance between spins go to zero. In this way, our discrete lattice becomes a continuous field, and we can reformulate the Ising model as a quantum field theory (QFT). Whereas the above are examples of renormalization in real-space, in QFT, we typically perform renormalization in momentum-space. Here, we can use Wilson's $4 - \epsilon$ expansion, treating the dimensionality of our system as a perturbation [18]. Then, with perturbation theory and diagrammatics, we approximate the values of our scaling variables.

### Monte Carlo Simulation

RG allows us to take advantage of the finite-size effects that dominate Monte Carlo techniques, and we can predict how our results will deviate from the infinite-size limits (see Appendix B.1). We can combine the two approaches, performing the RG sums (like equation (2.21)) over Monte Carlo samples.

### Kardanoff's Variational Technique

Another approximation scheme is that of Kadanoff. He writes the renormalization transformation as:

$$e^{-H'(s')} = \sum_s e^{\boldsymbol{T}_\lambda(s',s)-H(s)}, \tag{2.29}$$

with $T_\lambda$ serves as a function that couples the original and coarse-grained systems. Kadanoff derives upper- and lower-bound estimates of the free energy that depend on $T_\lambda$. By choosing $\lambda$ to tighten the bounds, Kadanoff variationally minimizes the difference in free energy between the initial and coarse-grained systems, $\Delta F = F(H'\ ) - F(H)$ (knowing the margin of error). However, this technique does not guarantee reasonable estimates of macroparameters. As Kadanoff himself observed:

> Hopefully, one might obtain good results for physical quantities by choosing the upper (lower) bound recursions that give the smallest error in the free energy. . . We say "hopefully" because usually one is not interested in the free energy itself. Rather its derivatives are of the major physical interest. Since the variational principles pertain to the free energy, there is no guarantee that the derivatives will be accurate [19].

This reflects one of the major challenges with RG. It is by no means "easy" to construct adequate RG schemes, and findings are rarely backed with rigorous justification. The details of appropriate transformations depend on the systems under investigation and require an amount of intuition on the part of the physicists [5]. We will see that, ultimately, being more precise in defining physically-relevant information will allow us to circumvent this problem. We will be able to formulate a system-independent criterion of quality for RG transformations.

## III. Machine Learning in Physical Investigations

In the previous chapter, we began with the goal of translating microscopics to macroscopics. Where MFT failed, RG made sense of critical phenomena, revealing universality, simple scaling laws, critical exponents, and the relations between them. In this chapter, we turn to machine learning. Like RG, ML is a blanket term that refers to a wide range of techniques. Both involve the iterative manipulation of information with the goal of extracting "relevant" information. However, we will see that ML is more flexible in its definition of relevant. Whereas in statistical physics relevant is synonymous with long-distance, in ML, relevance will depend on the problem at hand. We will also see that the level at which RG and ML manipulate information is different. Where statistical physics works with partition functions, ML works with probabilities. In practice, many of the challenges (namely, intractable sums) and the solutions are the same. In spirit, however, this distinction reveals something fundamental about the types of challenges that characterize statistical physics and ML.

In physics as a whole, our investigations are largely reductionist: we *begin* with a Hamiltonian which, in turn, defines a partition function, and our aim is to predict something about large sets of microstates. To this end, computational techniques, like MCMC algorithms, may use $P(\boldsymbol{s})$ (more precisely, the relative probability) to generate a finite set of samples $\mathcal{S}_{data}$ that, we hope, is representative of $\mathcal{S}$, the true, complete set of microstates. By this, we mean that the statistical properties of $\mathcal{S}_{data}$ and $\mathcal{S}$ should converge as we increase the number of samples. In ML, the investigation often works in the opposite direction: we are given some $\mathcal{S}_{data}$ and assume it is representative of $\mathcal{S}$. Then, our goal is to learn something about $P(\boldsymbol{s})$. While physics concerns properties of $\mathcal{S}$, ML concerns properties of $\boldsymbol{s}$. These are not absolute distinctions, but it will be valuable to bear them in mind.

Let us introduce some of the essentials of ML and show its value for statistical physics through a practical example. We follow the treatment of Hinton [20] as well as Mehta et al.'s *A high-bias, low-variance introduction to Machine Learning for physicists* [21]. We refer the interested reader to these sources for elaboration where our analysis goes quickly.

### Phase Classification

Our first question, as physicists, is the following: can we train a neural network to classify the likely phase of some Ising configuration, $\boldsymbol{x}$? This immediately reflects the different spirits of ML and physics we discussed above: our goal is to learn something about the microstate $s$, whereas in the previous chapter we cared only for $\mathcal{S}$. Formally, our aim is to learn the conditional probability distribution $P(\boldsymbol{y}|\boldsymbol{x})$, where $\boldsymbol{y}$ encodes the phase.

We start with a dataset. In this case, we assume that we have access to a large collection of Ising configurations at different temperatures and phases, $\mathcal{X}_{data} \coloneqq \{\boldsymbol{x}\}$, as well as their phase labels $\mathcal{Y}_{data} \coloneqq \{\boldsymbol{y}\}$. We often distinguish two kinds of ML: *supervised* and *unsupervised* learning.[8] Our access to labels, $\mathcal{Y}$, means this example of classification falls under supervised learning.[9] In contrast, unsupervised learning works without labels, typically at the level of $P(\boldsymbol{x})$, detecting patterns in the raw data itself.

Before doing anything else, we randomly divide $\mathcal{X}_{data}$'s elements into a training set, $\mathcal{X}_{train}$, from which our network learns, and a testing set, $\mathcal{X}_{test}$, on which we test the trained model's results. This division is crucial because, in ML, we want our results to generalize to new samples that may not have been in our dataset $\mathcal{X}_{data}$ but that could have come from $\mathcal{X}_{true}$. A central problem in ML is the *bias-variance tradeoff*. An algorithm may *overfit* the training set which compromises its ability to generalize to new data, or it might fail to learn enough detail, performing poorly on both training and testing data (low-bias/high-variance and high-bias/low-variance, respectively). By splitting the dataset into a testing and training set, we get a good impression of the bias and variance for the models we trained. Often, we will further split the training set into cross-validation sets, trying out different kinds of models and ultimately keeping those that best accomplish a balance of low bias and variance.

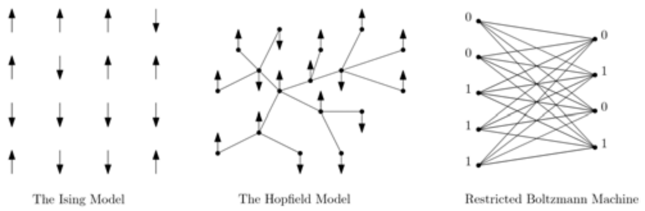


*Figure 3.1 – Restricted Boltzmann machines are a binary-valued two-layer neural network. RBMs trace their origins to the Hopfield model, an early model for associative memory which itself was inspired by the Ising model [22].*

A full review of the techniques available in ML is beyond the scope of this capstone. As such, we will focus our attention on one particular class of algorithms, those of restricted Boltzmann machines (RBMs): bidirectional artificial neural networks for modeling probability distributions. Depicted in Figure 3.1, RBMs consist of two layers of binary units: a *visible* layer, $\boldsymbol{v} \coloneqq \{v_i\}_{i=0}^{N}$, and a *hidden* layer, $\boldsymbol{h} \coloneqq \{h_j\}_{j=0}^{N}$, where $v_i, h_j \in \{0, 1\}$. The visible layer will serve as the input for our data, $\boldsymbol{x} \to \boldsymbol{v}$, and the hidden layer as our prediction, $\boldsymbol{h} \to \boldsymbol{y}$, the label. For the purposes of classifying the phases of the 2D Ising model, it will suffice to consider an RBM with one hidden unit, $h$. If $h = 1$, we predict that the configuration is in the ordered phase, and for $h = 0$, we predict it is disordered. From our exploration in the previous chapter, we could probably come up with some function that performs this prediction. For example, we might sum all of the spins (compute the magnetization) and if the result is close to $0$ we know the system is disordered: $h \coloneqq 0$. Otherwise, we would output $h \coloneqq 0$. The power of ML will be to extract rules like these without explicit instructions.

Instead, we teach our model implicitly through the cost function, $\mathcal{C}(P_\theta(y|\boldsymbol{x}), \{\mathcal{X}_{data}, \mathcal{Y}_{data}\})$. This acts as a moderator, and it tells the model how "incorrect" its predictions (labels) are. Formally, the ML goal becomes to find the parameters $\theta$ that minimize $\mathcal{C}$. Experience from theoretical physics tells us that finding the ground state (global minima) can be really, *really* hard, so instead, we use stochastic gradient descent (SGD). For each $\theta$, we calculate $\partial_\theta \mathcal{C}$, with which we implement the update rule:

[8] A third axis of differentiation, reinforcement learning is outside the scope of this capstone.

[9] We could, however, also formulate classification of phases as an unsupervised learning problem. We might try "clustering," where we specify a number of clusters, $k$, and try to group training examples into as many clusters, according to some notion of similarity. Then, we hope that the clusters our model learns correspond to the phases. We could also use this procedure with different choices of $k$ to determine a likely number of phases.

$$\theta \to \theta - \eta \partial_\theta \mathcal{C}, \tag{3.1}$$

where $\eta$ is the *learning rate*. From calculus, we know that the negative gradient of a function is the direction in which that function decreases most rapidly. This update procedure, then, adjusts our parameters so as to reduce the cost. Iterating this procedure, we end up in a local minimum of the cost function (see Figure 3.2). In practice, we calculate these gradients not over the entire data set, but over subsets – *minibatches*. This means the gradients will vary from iteration to iteration (hence, *stochastic*). This is both computationally faster and introduces noise (like temperature) that improves the chances of escaping poor local minima.

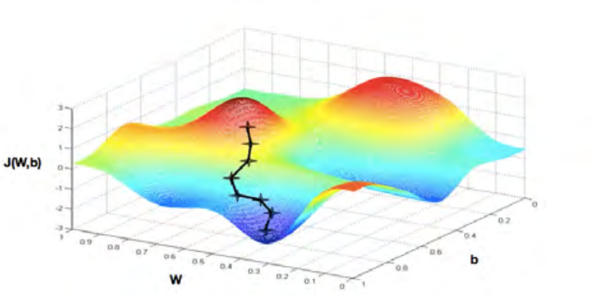


*Figure 3.2 – Stochastic Gradient Descent (SGD). If we view the cost function as defining an "energy" landscape, SGD allows us to find local minima (stable or metastable) energy states. Image taken from [23].*

Now, if we can formulate an adequate prediction function, $P(\boldsymbol{h}|\boldsymbol{v})$, for the RBM, we can improve it with SGD. The crucial step is to define an energy function:

$$\boxed{E_\theta(\boldsymbol{v},\boldsymbol{h}) := -\sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} w_{ij} v_i h_j} \tag{3.2}$$

where $\theta := \{\{a_i\},\{b_j\},\{w_{ij}\}\}$. Then, we can model the system's joint probability with a Boltzmann distribution:

$$\boxed{P_\theta(\boldsymbol{v},\boldsymbol{h}) := \frac{1}{Z} e^{-E_\theta(\boldsymbol{v},\boldsymbol{h})}, \qquad Z := \sum_{\boldsymbol{v}',\boldsymbol{h}'} e^{E_\theta(\boldsymbol{v}',\boldsymbol{h}')}}$$

(3.3)

We note that this is the exact same equation as our Ising model (2.9 and 2.10). Here, $\{a_i\}$ and $\{b_j\}$ take the role of the external magnetic field $B$, which now varies from site to site (hence the indices $i$ and$j$). Then $\{w_{ij}\}$ takes the role of $J$ varying from pair to pair.

Naturally, we run into the same intractability issues. For large enough networks we cannot evaluate $Z$. Instead of calculating joint distributions, we, similar to MFT and MCMC, consider instead conditional and marginal distributions. Due to RBM's bipartite structure these factors are easy to evaluate. Explicitly, in the case of a single hidden spin $h_j$, we get (with a bit of algebra):

$$\begin{aligned} P(h_j|\boldsymbol{v}) = \frac{P(h_j,\boldsymbol{v})}{P(\boldsymbol{v})} &= \frac{(e^{-E(\boldsymbol{v},h_j)})/Z}{(\sum_{\boldsymbol{h}'} e^{-E(\boldsymbol{v},\boldsymbol{h}')})/Z} \\ &= \frac{e^{-h_j(\sum_i w_{ij} v_i + b_j)}}{1 + e^{(\sum_i w_{ij} v_i + b_j)}}. \end{aligned}$$

(3.4 & 3.5, respectively)

This conditional probability is itself a Boltzmann distribution, but over only two states (tractable!). We can use this to express to the full system as:

$$\boxed{P_\theta(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^{M} \frac{1}{1 + e^{-h_j(\sum_i w_{ij} v_i + b_j)}}} \tag{3.6}$$

and similarly:

$$\boxed{P_\theta(\boldsymbol{v}|\boldsymbol{h}) = \prod_{i=1}^{N} \frac{1}{1 + e^{-v_i(\sum_j w_{ij} h_j + a_i)}}} \tag{3.7}$$

All that remains is for us to choose an adequate cost-function, and we can start training our RBMs. For the example of binary-classification, an appropriate choice is the cross-entropy loss:

$$\mathcal{C}(P_\theta(\boldsymbol{y}|\boldsymbol{x}), \{\mathcal{X}_{data}, \mathcal{Y}_{data}\}) = \sum_{\boldsymbol{x}\in\mathcal{X}_{data}} \sum_{\boldsymbol{y}\in\{0,1\}} P_{true}(\boldsymbol{y}|\boldsymbol{x}) \log P_\theta(\boldsymbol{y} \tag{3.8}$$

see Appendix C.1 for elaboration. Differentiating with respect to $\theta$ and with some simple algebra, we get the training rule:

$$\partial_{b_j}\mathcal{C} = \sum_{\boldsymbol{x}\in\mathcal{X}_{batch}} P_{true}(\boldsymbol{y}|\boldsymbol{x})\big[1 - P_\theta(\boldsymbol{y}|\boldsymbol{x})\big]\,(h_j)\,,$$
$$\partial_{w_{ij}}\mathcal{C} = \sum_{\boldsymbol{x}\in\mathcal{X}_{batch}} P_{true}(\boldsymbol{y}|\boldsymbol{x})\big[1 - P_\theta(\boldsymbol{y}|\boldsymbol{x})\big]\,(v_i h_j)\,.$$

(3.9 & 3.10, respectively)

We have all the necessary elements of a phase classifier: a dataset, a model of $P(\boldsymbol{y}|\boldsymbol{x})$, and a means to train this model. In the next section, we will consider an example more relevant to the statistical physicist: generating samples.

## Generative Modelings: Gibbs Sampling

In Section 2.3, we considered how to use MCMC sampling to estimate macroparameters and even critical exponents. Here we will consider an MCMC technique called *Gibbs Sampling* which will use RBMs to generate samples of $P(\boldsymbol{x})$.

Suppose we are given an already trained RBM. Gibbs sampling, like other MCMC techniques, consists of a series of update steps. In one step, we input some initial state, transform it into a hidden state using $P_\theta(\boldsymbol{h}|\boldsymbol{v})$, and transform it back to a new visible state using $P_\theta(\boldsymbol{v}|\boldsymbol{h})$. This process is imperfect, so the output of one step will differ from the input, and this difference will become increasingly smaller as the network learns. If we repeat this process many times, then the distribution of the outputs will converge to the equilibrium distribution $P_\theta(\boldsymbol{v})$..

This first requires $P_\theta(\boldsymbol{v})$ to be an appropriate model of $P_{true}(\boldsymbol{v})$. To get to this point, we need to derive a marginal distribution over $\boldsymbol{v}$ from $P_\theta(\boldsymbol{v},\boldsymbol{h})$. Similar to the conditional probabilities, the architecture of the RBM allows us to perform the marginalization $P_\theta(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P_\theta(\boldsymbol{v},\boldsymbol{h})$ explicitly. If we write $P_\theta(\boldsymbol{v})$ as a Boltzmann distribution with its own energy $E_\theta(\boldsymbol{v})$,

$$P_\theta(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P_\theta(\boldsymbol{v},\boldsymbol{h}) \propto e^{-E_\theta(\boldsymbol{v})},$$

(3.11)

then we can express $E_\theta(\boldsymbol{v})$ in terms of $E_\theta(\boldsymbol{v},\boldsymbol{h})$ (see Appendix D.1):

$$\boxed{E_\theta(\boldsymbol{v}) = -\sum_i a_i v_i - \sum_j \log\left(1 + \exp\left\{-\left(b_j + \sum_i v_i w_{ij}\right)\right\}\right)}$$

(3.12)

and we can perform an analogous computation for the marginal distribution over $\boldsymbol{h}$ to get $P_\theta(h)$ as a Boltzmann distribution in terms of energy, $E_\theta(h)$. As in the classification example, we require a cost function – in this case, the Kullback-Leibler divergence (KLD):

$$\mathcal{C}(P_\theta(\boldsymbol{x}),\boldsymbol{x}) = D_{KL}(P_{data}(\boldsymbol{x})||P_\theta(\boldsymbol{x})) \coloneqq \sum_{\boldsymbol{x}\in\mathcal{X}_{data}} P_{true}(\boldsymbol{x}) \ln \Bigg( \frac{P_t}{P}$$

(3.13)

This is closely related to the cross-entropy and, similarly, it is an important information-theoretic quantity (see Appendix C.2) that provides a notion of similarity for probability distributions. It is $0$ if and only if the two distributions are equal:

$$D_{KL}(P_{true}(\boldsymbol{x})||P_\theta(\boldsymbol{x})) = 0 \iff P_{true}(\boldsymbol{x}) = P_\theta(\boldsymbol{x}).$$

(3.14)

We claimed that this is unsupervised, but in comparing $\mathcal{C}$ in equation (3.13) with equation (3.8), we might interpret $\boldsymbol{x}$ as a kind of label for itself. A more appropriate description is *self-supervised*. We emphasize this point because it has been a source of misunderstanding. ML always requires the specification of a cost function, and even in the absence of labels, the cost function guides how the model learns which information is relevant and irrelevant.

If we differentiate the KLD with respect to $\theta$, then after a bit of algebra, we get:

$$\partial_{a_i} D_{KL} = \langle v_i\rangle_{true} - \langle v_i\rangle_\theta$$
$$\partial_{b_j} D_{KL} = \langle h_j\rangle_{true} - \langle h_j\rangle_\theta$$
$$\partial_{w_{ij}} D_{KL} = \langle v_i h_j\rangle_{true} - \langle v_i h_j\rangle_\theta$$

(3.15, 3.16 & 3.17, respectively)

where $\langle\ldots\rangle_{true}$ is an average over the true distribution $P_{true}(\boldsymbol{v})$ and $\langle\ldots\rangle_\theta$ is an average over the model distribution $P_\theta(\boldsymbol{v})$.

Naturally, we approximate. Wherever $h_j$ shows up, we use $P(h_j|\boldsymbol{v})$, and for the expectations over $P_{true}$, we calculate a Monte Carlo average over our dataset, $\mathcal{X}_{data}$. The expectations over $P_\theta$ are trickier. Fortunately, we can approximate them with Gibbs sampling, initializing with samples from $\mathcal{X}_{data}$. Together, these approximations constitute the *contrastive-divergence* algorithm, see [20]. Having trained our RBM, we can generate new samples and start calculating critical exponents.

RBMs trained in this way have applications other than generation. Consider training these on black and white images: if we constrain the number of hidden nodes to be less than the number of visible nodes, then the hidden layer will learn a compact representation of the input. We can even stack multiple RBMs on top of each other to form *deep Boltzmann machines* (DBMs). In DBMs, the hidden layer of one RBM becomes the visible layer of the next. Trained with contrastive divergence, each layer learns a progressively more compact version of the input. This should remind you of RG. In the next section, we will make these similarities more exact. Ultimately, this allows us to machine learn RG transformations.

## IV. Machine Learning and the Renormalization Group

In the previous chapter, we considered two examples of ML in physical investigations: neural networks used as a phase classifier and MCMC sampler. Consider the superficial similarities between these neural networks and our treatment of RG in Chapter 2. The classification example calls to mind the infinite RG limit in which all disordered phases flow to one fixed point and all ordered phases to another. Here, these fixed points would correspond to the values of our label being either $0$ or $1$. The generative example is more immediately similar, and the very language is analogous: we speak of hidden layers that hierarchically decompose the input visible layer to coarse-grained hidden layers. We might be tempted to ask: does a DBM of, say, five layers learn to implement four iterations of some RG procedure (see Figure 4.1)?

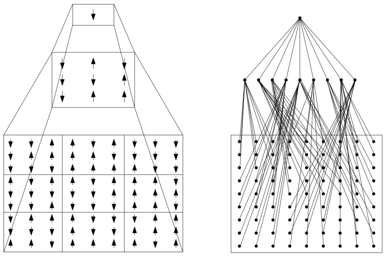


*Figure 4.1 – Two iterations of block renormalization and a deep Boltzmann machine of three layers.*

However, questions like this are too general. We have already seen two examples of RBMs used for different goals, and what constituted relevant information differed in either context. In RG, relevant information is understood more narrowly: it is the long-distance information. If we are to make comparisons between RG and RBMs, we must be more specific, and this begins by being more precise in our definition of 'relevant' information. Although Claude Shannon, the founder of information theory, avoided this topic explicitly, Tishby et al. showed that information theory provides a natural and exact formulation of relevant information: relevant information is simply the information contained in one signal, $\boldsymbol{x}$, about another $\boldsymbol{y}$ [3]. For our phase classifier, the relevant information is the information contained in the Ising samples $\boldsymbol{x}$ about the labels $\boldsymbol{y}$. Once we have trained RBMs for compression, the relevant information is the information contained in the hidden layer $\boldsymbol{h}$ about the visible layer $\boldsymbol{v}$. Information theory quantifies how much information is shared between two signals with the mutual information:

$$I(\boldsymbol{x};\boldsymbol{y}) := \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{P(\boldsymbol{x},\boldsymbol{y})}{P(\boldsymbol{x})P(\boldsymbol{y})}\right), \tag{4.1}$$

We see this quantity is minimized ($I(\boldsymbol{x};\boldsymbol{y}) = 0$) when the random variables are independent ($P(\boldsymbol{x},\boldsymbol{y}) = P(\boldsymbol{x})P(\boldsymbol{y})$). It is maximized when the variables are maximally dependent. We have a quantitative basis for extracting relevant information: maximizing mutual information between appropriately chosen signals.

In the next chapter, we will derive the correct choice for $\boldsymbol{x}$ and $\boldsymbol{y}$ on lattice systems that recovers physically-relevant (i.e. long-distance) information. We then derive a variational proxy to the mutual information that we can differentiate and, thus, learn in a neural network.

With the knowledge of an exact formulation of relevant information, we can start investigating links between particular RG implementations and RBMs. Such comparisons will revolve around the question of relevance. What information does a given cost-function deem relevant? Are there examples where this overlaps with our notions of physical relevance? Let us consider a seminal example: the equivalence posed by Mehta and Schwab.

## A Comment on Mehta and Schwab's Equivalence

As a first attempt, we will consider Mehta and Schwab's correspondence between Kadanoff's variational procedure (Chapter 2.3) with RBMs trained by contrastive divergence (Chapter 3.2) [4]. Specifically, we consider the narrower case, of "exact" RG and "exact" RBMs. Exact RG means that Kadanoff's transformation preserves *all* the information contained in the hidden system; i.e. the free energy of the input and coarse-grained systems is exactly the same. Similarly, an exact RBM has learnt to perfectly recreate its inputs $P_{true}(\boldsymbol{x}) = P_\theta(\boldsymbol{x})$ in equation 3.13; the RBM will have reached a global minimum of the cost function.

First, Mehta and Schwab show that, under the above conditions, the two transformations are equivalent under:

$$\boldsymbol{T}_\theta(\boldsymbol{v}, \boldsymbol{h}) = -\boldsymbol{E}_\theta(\boldsymbol{v}, \boldsymbol{h}) + \boldsymbol{H}(\boldsymbol{v}), \tag{4.2}$$

and we provide the derivation in Appendix D.2.

However, the exact case is not generally possible, and when it is, it is often not particularly interesting. Exact RBM transformations likely mean overfitting which is often opposed to the aims of the ML investigation. Exact RG transformations are few and far between, so this correspondence would apply only under narrow circumstances. More interesting then, would be to consider a non-exact comparison.

Here, the two approaches will vary: Kadanoff's variational method operates at the level of free energies while contrastive divergence works at the level of probabilities. Optimizations at these different levels will not necessarily coincide. Still, there may be qualitative similarities. Pursuing this line of inquiry, Mehta and Schwab train RBMs as described in Chapter 3.2. They find that the RBMs learn to couple hidden units to local blocks of visible units. Crucially, this was not possible without *regularization*, a technique that promotes sparse connections, between hidden and visible units see [21].

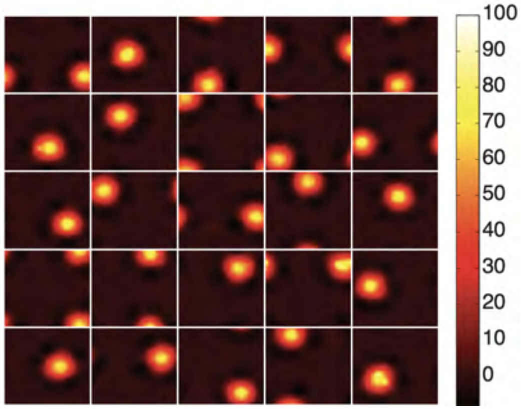


*Figure 4.2 – Visualization of the receptive filters of the hidden units in one of the layers of the DBM trained by Mehta and Schwab [4]. Each plot corresponds to one hidden unit, and each pixel corresponds to one spin in the initial layer. Higher intensities mean the hidden unit is more coupled to that spin. Indeed, we see that the RBM learns to couple local transformations. However, the transformations do not preserve nearest neighbor relations. This does not matter for the total amount of information stored in coarse-grained layers. In RG, these concerns do matter because they affect the practicality and interpretability of the result [5].*

Mehta and Schwab write:

> Surprisingly, this local block spin structure emerges from the training process, suggesting the [deep neural network] is self-organizing to implement block spin renormalization [4].

These RBMs may indeed learn a kind of block-spin transformation. However, this block-spin transformation need not be block-spin *renormalization*. As Koch-Janusz and Ringel put it:

> [T]he usefulness (and practicality) of the RG procedure depends on choosing [the transformation] . . . such that the effective Hamiltonian. . . remains as short range as possible [5].

More precisely, in the Taylor expansion of our coarse-grain Hamiltonians, the shortest range terms should dominate:

$$H(\boldsymbol{s}) = -\sum_i K_i^{(1)} s_i - \sum_{\langle i,j \rangle} K_{ij}^{(2)} s_i s_j - \sum_{\langle\langle i,j \rangle\rangle} K_{ij}^{(3)} s_i s_j \dots, \tag{4.3}$$

where $K^{(n)}$ are exponentially suppressed in $n$. In RG, we place additional constraints on how to organize the information in subsequent layers. Block renormalization procedures, including Kadanoff's technique, respect the symmetries and topology of the system under consideration. The locality of interactions means neighboring blocks of visible spins become neighboring individual hidden spins. Translational symmetry means the same transformation is applied to each block. Mehta and Schwab's RBMs will not, in general, satisfy these two conditions.

Let us be more exact by rephrasing renormalization in probabilistic terms: we parametrize our RG transformation as the conditional probability distribution distribution $P(\boldsymbol{h}|\boldsymbol{x})$, where $\boldsymbol{x}$ is our initial configuration and $\boldsymbol{h}$ is the hidden or coarse-grained configuration. Locality and translational invariance of the Ising model mean we can factor the RG transformation into a product of local single-block transformations. We divide the system into $M$ blocks, $\boldsymbol{x} \to \{v^{(1)}, v^{(2)}, \dots, v^{(M)}\}$, with corresponding hidden units $\{h_1, h_2, \dots, h_M\}$. Then we can factor $P(\boldsymbol{h}|\boldsymbol{x})$ as:

$$P(\boldsymbol{h}|\boldsymbol{x}) = \prod_{j=1}^{M} P(h_j|\boldsymbol{v}^{(j)}). \tag{4.4}$$

Consider what happens, when you permute the units in the hidden layer; i.e. use a different transformation transformation $P(h_{j'}|\boldsymbol{v}^{(j)})$ where $j \neq j'$. As long as our reverse rule, $P(\boldsymbol{x}|\boldsymbol{h})$ also takes this into account, such a permutation has no impact on the performance of Mehta and Schwab's contrastive divergence trained RBMs. This is *not* the case in RG. There is a unique permutation of hidden layer degrees of freedom which maximizes the above short range condition, and acceptable RG transformations identify this permutation.

Furthermore, the block transformations that the RBM learns may not be the same for each block. Compression and generation are invariant under partial flips $h_j : (0,1) \to (1,0)$ for some fixed $j$.[10] Blocks of spins which were perfectly correlated in the visible layer may be perfectly anticorrelated in the hidden layer. Although the information content is the same, the hidden layer Hamiltonian will take a more complicated form than is necessary, going against notions of what is a good and practical RG procedure.

Mehta and Schwab fail to mention these conditions, so their suggestion that RBMs learn block-spin renormalization falls short. However, if we make these conditions explicit, we can easily recover a stronger version of their claim. To do this, we introduce convolutional neural networks (CNNs) as depicted in Figure 4.3. These are networks with a special architecture: instead of all-to-all couplings, we explicitly couple spins in the hidden layer to blocks in the visible layer. Additionally, CNNs perform the same translation for each block and respect the topology of the input system. Had they required a convolutional architecture upfront, Mehta and Schwab may, indeed, have successfully trained RBMs to learn even block renormalization.

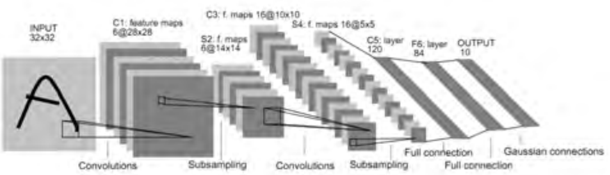


*Figure 4.3 – This is LeNet-5, an example of a convolutional neural network used for digit recognition and a seminal architecture [24].*

From these observations, we can make the link between majority-rule block-spin RG and ML fully precise. In Appendix D.3, we derive an equivalence between RBMs and RG, describing a parameterization which implements the majority-rule: $P(h_j|\boldsymbol{v}^{(j)}) = \operatorname{sgn} \sum_i v_i^{(j)}$.

## Relevant Information

For the exact case discussed by Mehta and Schwab, we preserve the full probability measure which necessarily preserves the long-distance information. However, in the non-exact case, the KLD has no preference for long-distance information over short-distance: all information is valued equally.

[10] RG transformations are invariant under a collective flip of $\boldsymbol{h}$.

This is acceptable in the case of compression and generation, as we have no a priori knowledge of which features in the data are more important.

In RG, we have stronger requirements: transformations must favor long-distance information over short-distance information. We can say resolutely: RBMs trained with the KLD (even with the appropriate convolutional architecture) need not learn acceptable RG transformations. Perhaps in the case of the simple Ising model, the KLD is an adequate heuristic. However, we have no guarantee that this extends to novel systems. In fact, Koch-Janusz and Ringel showed that RBMs trained with the KLD on the dimer model, another model from statistical physics, couple to local fluctuations rather than to the correct hidden variables [5]. We recall that appropriate transformations for one system need not translate to others (see Chapter 2.3). Critique of Mehta and Schwab's paper revolves primarily around the KLD being an inappropriate criterion for RG transformations [2, 6].

In reply, Mehta and Schwab point out that Kadanoff's method also offers no guarantee of extracting the correct physical information. We come back to the difficulty of devising appropriate RG techniques, and the need for creativity. In fact, this need for intuition in RG is mirrored quite clearly in ML. Decisions regarding model architectures and cost functions require creativity on the part of the researcher. Both fields would benefit from a clearer understanding of when techniques will and will not work. In the next chapter, we will see that information theory may provide the framework to answer questions like these, and we will encounter a system-independent formulation of RG. This might mean better, more efficient models in both disciplines.

# V. Renormalization in Information Theory

In the previous chapter, we used information theory to formalize the notion of *relevant* information. In this chapter, we will adapt this to our physical context of long-distance as relevant information. Ultimately, this allows us to devise a cost-function that measures the physical information. From this, we can derive a learning rule to train RBMs to perform *optimal* transformations.

In the previous chapter, we saw that the Ising model's locality and translational invariance conditions reduce the number of acceptable RG procedures. Rather than devise an update rule for how an entire configuration should transform under RG, we can start with equation (4.4). Then, we only need to learn the transformation for a single block, $P(h_j|\boldsymbol{v}^{(j)})$. We already know that we can model conditional probabilities like these with RBMs. This is the first insight of Koch-Janusz and Ringel, and in this chapter we will follow their information-theoretic treatment [5].

## An Information-Theoretic Formulation of Renormalization Group

We first partition a full lattice configuration $\boldsymbol{x}$ into four areas: a visible block, $\boldsymbol{v}$, that is surrounded by, in order, a buffer, $\boldsymbol{b}$, an environment, $\boldsymbol{e}$, and an outer area, $\boldsymbol{o}$ (see Figure 5.1). In RG, we introduce a hidden area, $\boldsymbol{h} \coloneqq \{h_i\}$ which is a function of the degrees of freedom in $\boldsymbol{v}$. Our aim is to choose this coupling, $P(\boldsymbol{h}|\boldsymbol{v})$, so that $\boldsymbol{h}$ encodes the long-distance degrees of freedom about our system at large, $\boldsymbol{x}$[11]. By our assumption that $\boldsymbol{o}$ is farther than the correlation length, $\boldsymbol{v}$ contains no information about $\boldsymbol{o}$. We can ignore $\boldsymbol{o}$ in devising $P(\boldsymbol{h}|\boldsymbol{v})$. Furthermore, in coming up with a function for $\boldsymbol{h}$, we can reasonably ignore $\boldsymbol{b}$. The information contained in $\boldsymbol{v}$ about $\boldsymbol{b}$ is likely to be short-range. By eliminating $\boldsymbol{b}$, we effectively remove the shortest-range fluctuations.

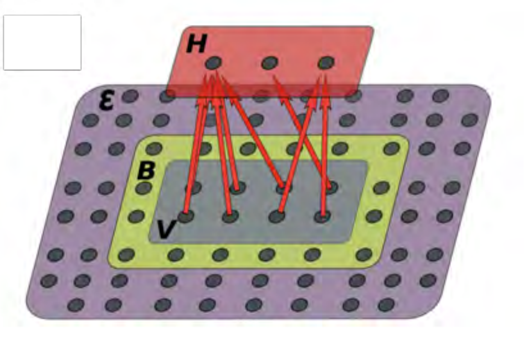


*Figure 5.1 – The RSMI algorithm partitions configurations into four regions, a visible block* $\boldsymbol{v}$*, buffer zone* $\boldsymbol{b}$*, environment* $\boldsymbol{e}$ *and outer zone,* $\boldsymbol{o}$*. We introduce a coarse-grained block of spins* $\boldsymbol{h}$*.*

That which remains is the physically relevant information; we see it is the information shared between $\boldsymbol{v}$ and $\boldsymbol{e}$. Our goal is to extract this information

[11] We have dropped the indices for convenience, and now we let block h consist of multiple coarse-grained spins.

and encode in $\boldsymbol{h}$. We choose the parameters, $\Lambda$, of our RBM that models $P(\boldsymbol{h}|\boldsymbol{v})$ to maximize the mutual information between $\boldsymbol{h}$ and $\boldsymbol{e}$:

$$I(\boldsymbol{h};\boldsymbol{e}) = \sum_{\boldsymbol{h},\boldsymbol{e}} P(\boldsymbol{h},\boldsymbol{e}) \log\left(\frac{P(\boldsymbol{h},\boldsymbol{e})}{P(\boldsymbol{h})P(\boldsymbol{e})}\right), \tag{5.1}$$

where these probabilities are defined as marginalizations over $P(\boldsymbol{x})$:

$$\begin{aligned} P(\boldsymbol{h},\boldsymbol{e}) &= \sum_{\boldsymbol{v}} P(\boldsymbol{h}|\boldsymbol{v})P(\boldsymbol{v},\boldsymbol{e}) \\ P(\boldsymbol{h}) &= \sum_{\boldsymbol{v},\boldsymbol{e}} P(\boldsymbol{h}|\boldsymbol{v})P(\boldsymbol{v},\boldsymbol{e}) \\ P(\boldsymbol{v},\boldsymbol{e}) &= \sum_{\boldsymbol{b},\boldsymbol{o}} P(\boldsymbol{x}) \\ P(\boldsymbol{e}) &= \sum_{\boldsymbol{b},\boldsymbol{v},\boldsymbol{o}} P(\boldsymbol{x}). \end{aligned}$$

(5.2, 5.3, 5.4 & 5.5, respectively)

From previous chapters, we know that the probability measure $P(\boldsymbol{x})$ and its marginalizations are generally intractable. The key insight of Koch-Janusz and Ringel is that we can use RBMs not only to model $P(\boldsymbol{h}|\boldsymbol{v})$ but also to model these other distributions. Koch-Janusz and Ringel ultimately use three RBMs: one $P(\boldsymbol{v},\boldsymbol{e})$, another for $P(\boldsymbol{v})$, and finally the $P(\boldsymbol{h}|\boldsymbol{v})$ RBM already mentioned. The other distributions are calculated as MC-averages over the dataset. Hereby, Koch-Janusz and Ringel derive a proxy to the mutual information (see Appendix E.1) which they can optimize with stochastic gradient descent.

To validate their ideas, they provide both experimental and theoretical justification. For the 1D Ising model, these marginalizations can be calculated exactly, and they show that maximizing the RSMI rederives decimation, an RG procedure known to be *optimal* for the 1D Ising model in that the procedure does not increase the range of the coarse-grained Hamiltonian. In follow-up research, Lenggenhager et al. (along with the aforementioned authors) show that RSMI coarse-graining is more generally optimal, maintaining short-distance in any number of dimensions. They further show this holds even in some cases that the mutual information is not fully saturated. We direct the interested reader to [5] and [9] for these results. In the remainder of this capstone, we discuss our own implementation and generalization.

### Measuring Critical Exponents

Having trained an RBM according to the RSMI algorithm, we can use finite-size scaling (see Appendix B.1) to estimate, finally, critical exponents. An important detail will be to devise a “thermometer” that can measure the effective temperature at successive RG-iterations. We discuss several options in Appendix E.2. These options, moreover, are intrinsic and do not require explicit knowledge of the temperature. This means the RSMI algorithm is fully unsupervised. If we encounter new systems where we do not know the proper RG transformations (or even how to measure “temperature”), then unsupervised approaches might save us considerable headache. Rather than guess and check possible transformations, we would first train a neural network to learn how to perform a renormalization procedure on the system, then observe and interpret, only later attempting more calculation-heavy approaches. As Koch-Janusz and Ringel put it, the RSMI algorithm could form the basis of the “physical reasoning process” itself kjr]. Instead of retroactively explaining trends in our data, ML would proactively guide our exploration.

## VI. Results: Machine Learning Critical Exponents

In [10], we release *rgpy*, an open-source library for implementing ML-based renormalization group techniques. This is still in its infancy, and development is ongoing. As of now, it contains a full-stack realization of the RSMI algorithm implemented in Tensorflow: i.e., the package includes implementations of various MCMC techniques (Metropolis-Hastings, Swendsen-Wang, and Wolff algorithms), the RBMs necessary for the RSMI algorithm, and an implementation of standard block-spin renormalization. We also provide a host of already-generated samples at various lattice sizes and temperatures. We invite the reader to explore and try out these tools.

## A Novel Calculation of $\nu$ for the 2D Ising Model

Our exploration of statistical physics and ML was centered around the Ising model and its macroparameters and critical exponents. The calculation of these exponents served as unifying thread, and as a validation of the RSMI algorithm we provide the following approximation of $\nu$:

$$\boxed{\nu \approx 0.79 \pm 0.39} \tag{6.1}$$

For how we calculated this, see Appendix E.3. This is not, by any means, an improved calculation. However, it is a highly promising first result. The quality of this calculation was ultimately limited by both time and hardware constraints. With more time and resources, the prediction should further converge to the right quantity.

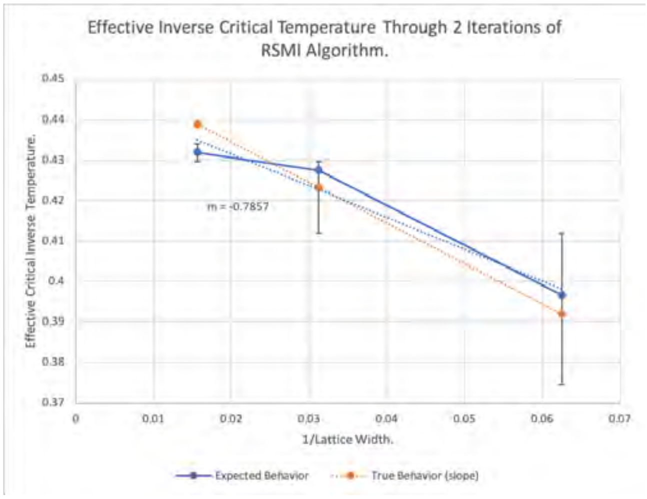


*Figure 6.1 – The finite-size scaling curve for the correlation length critical exponent.*

In fact, this result is just the start. Our investigation revealed a large number of possible improvements and generalizations. Due to the time constraints of this capstone, we have not yet implemented these ideas, but all the same, they merit attention. The generalization we discuss gives rise to a family of RSMI-inspired techniques, together constituting a new branch of RG methods.

## A Generalization to $n$-Spin and $O(n)$ Systems

Koch-Janusz and Ringel claim the RSMI algorithm works for general lattice systems. However, the current formulation works only for systems with binary degrees of freedom: spin-$1/2$ Ising models. The authors avoid explicitly generalizing these results. Though the generalization is straightforward, it is important enough to warrant elaboration.

Let us consider systems with either $n$ spins or $n$ components of spin, then, the visible units of our RBM should be able to take $n$ values. Accomplishing this is quite simple: we let our visible units become vectors with $n$ components: $v_i \to \{v_{id}\}_{d=1}^n$. We say that $v_i$ takes the value of spin $d$ when:

$$v_i \equiv d \iff \begin{cases} v_{id} = 1 \\ v_{id'} = 0, d' \neq d. \end{cases} \tag{6.2}$$

In fact, we will use the RBM to model a slightly different vector:

$$v_{id} = P(d), \tag{6.3}$$

where the vector is normalized $\sum_d v_{id} = 1$. Such a vector that contains probabilities of classes is called a *one-hot encoding*, and with these probabilities, we can randomly sample values of spin to generate $n$-valued spins. For $n$-vector models, we leave the probabilities as they are. To allow our RBMs to produce this encoding, we have to introduce the index $d$ in the parameters of our model:

$$\begin{aligned} a_i &\to a_{id} \\ w_{ij} &\to w_{ijd}. \end{aligned}$$

(6.4 & 6.5, respectively)

To derive a one-hot encoding, we change the RBM's reverse update rule, $P(\boldsymbol{v}|\boldsymbol{h})$ (3.7). The conditional probabilities will still factor, but normalization over the scalar values $v_i = 0, 1$, becomes normalization over the components of the vector, $d$. If we denote $\epsilon_i := -(\sum_j w_{ij} h_j + a_i)$, then, whereas before, we had:

$$P(v_i|\boldsymbol{h}) = \frac{e^{-\epsilon_i v_i}}{\sum_{v_i' \in \{0,1\}} e^{-\epsilon_i v_i'}}, \tag{6.6}$$

now we have, $\epsilon_{id} := -(\sum_j w_{ijd} h_j + a_{id})$, so:

$$P(v_{id}|\boldsymbol{h}) = \frac{e^{-\epsilon_{id} v_{id}}}{\sum_{d=1}^n e^{-\epsilon_{id} v_{id}'}}.$$

(6.7)

For $n = 2$, we re-derive the previous rule with a suitable choice in $w_{id}$, $a_{id}$. For reasonable values of $d$ (anything we might encounter in practice), we can perform these calculations explicitly. If we apply this change to each of the RBMs in the RSMI algorithm, we will be able to model $n$-spin and $O(n)$ models. Note that since we did nothing to the hidden variables, these will still be either binary or Bernoulli-valued. We may require that the hidden spins take the same form as the input variables, in which case we also need to apply a completely analogous change to the hidden unit of the $P(h_j|\boldsymbol{v}^{(j)})$ RBM. This is often the case in RG implementations, and it allows us to apply the RSMI algorithm to generic lattice systems.

There is, in fact, more room for generalization. Consider an arbitrary Hamiltonian for a system with binary-valued data $H(\boldsymbol{x})$. If we Taylor expand this function, we would get:

$$H(\boldsymbol{x}) = a + \sum_i b_i x_i + \sum_{ij} c_{ij} x_i x_j.$$

(6.8)

These are *all* the terms in the expansion.[12] Then, barring the bipartite structure, the RBM energy function already has the most general form we could possibly consider. However, when we make the switch to continuous or $n$-ary valued data, this is no longer true: the Taylor-expansions become infinite. In these cases, we might consider energy functions with different, possibly higher-order interactions. As long as we maintain the bipartite structure, we can ensure that conditional probabilities factor. Then, we can use variants of contrastive-divergence to efficiently train these models. We need not even require that $P(\boldsymbol{v}|\boldsymbol{h})$ and $P(\boldsymbol{h}|\boldsymbol{v})$ take the same form. The first might be Gaussian and the second a gamma distribution. Formally, we can consider the extension of RBMs into the *exponential family*, and for a full discussion of the possibilities, we refer the reader to [25]. This extension offers more powerful options for modeling probability distributions. Though the choices implemented in this capstone suffice for the spin-$1/2$ Ising model, it may be necessary to extend the work to the exponential family to investigating more complicated systems.

For all the exact results of Koch-Janusz and Ringel, their practical implementation uses a proxy which does not fully capture the above mutual information [5]. Here, we see two possibilities. First, we can improve Koch-Janusz and Ringel's approximation. The authors use a cumulant expansion

$$\langle \exp\{K(\boldsymbol{x})\}\rangle = e^{\sum_{\kappa=0}^{\infty} \frac{1}{\kappa!} C_\kappa},$$

(6.9)

with cumulants expressed in terms of the moments

$$\begin{aligned} C_1 &= \langle K \rangle, \\ C_2 &= \langle K^2 \rangle - \langle K \rangle^2, \\ C_3 &= \langle K^3 \rangle - 3\langle K^2 \rangle \langle K \rangle + 2\langle K \rangle^3, \end{aligned}$$

(6.10, 6.11 & 6.12, respectively)

and so on. In deriving the proxy, only the first term is kept (see Appendix E.1). For better results, a first improvement could be made by introducing additional terms.[13] Second, we can consider other approximations of the mutual information, and we draw the reader's attention to a sampling of options in the literature: contrastive predictive coding [26], Deep InfoMax [27], and Mutual Information Neural Estimation [28]. All three examples use mutual information as the basis for powerful unsupervised techniques. Though we avoid details, similarities (and differences) with the RSMI algorithm warrant further investigation.

Perhaps the most exciting future course of action is extending these ideas to momentum-space renormalization. Rather than focus on lattice models, our goal would be to solve quantum field theories. A first attempt might use the quantum-to-classical mapping of the path integral formalism [5]. More powerful however would be to use the same physical and information-theoretic arguments to devise conditions equivalent to equation (5.1) from first fundamentals.

## VII. Discussion and Conclusions

Let us summarize what we have accomplished. We began with an overview of the basic elements

[12] For $p > 1$, $x_i^p = 0$ or $1$.

[13] We may even be able to avoid this expansion altogether, see Appendix E.1.

of statistical physics, which has the goal of turning microscopic theories into concrete macroscopic observables. One quickly ran into a problem: intractable probabilities. To avoid this fundamental problem, we sampled a host of techniques. MCMC techniques avoid absolute probabilities with iterations of relative probabilities. Mean field theory uses a clever constraint on conditional distributions. RG creates new probability distributions, iteratively simpler, keeping the relevant long-distance information. After an introduction to ML, we noticed a resemblance between neural networks and RG. We identified a need for a more precise notion of 'relevant' information, which one formalizes with information theory. At the level of 'relevant' information, ML and RG appeared to behave similarly, extracting relevant information and suppressing irrelevant information. However, we saw that RG contained a more narrow conception of relevance: limited to long-distance.

We evaluated Mehta and Schwab's seminal comparison of Kadanoff's RG and CD-trained RBMs and identified a flaw in one of their claims: a 'good' RG should uncover compact, short-range hidden representations and respect the original system's symmetries. Evidently, RBMs do not obey these constraints these unless explicitly instructed to do so. By being more precise in formulating conditions, requiring convolutional architectures, we derived an exact correspondence between majority-rule block-spin renormalization and restricted Boltzmann machines.

Using the information-theoretic formulation of relevance, one can derive a system-independent RG criterion that *optimally* satisfies the short-distance condition. We implemented these ideas in *rgpy*, performing a recalculation of the critical length correlation exponent. This library is publically available, and we will continue developing it into the future. We encourage the reader to try this out, and we hope to enable many more calculations of critical exponents. To kickstart this project, we provided an overview of possible improvements and generalizations: the next likely targets are the XY-model and the spin-1 Ising model.

We place special emphasis on the role played by physical reasoning throughout our exploration. Though information theory presented us a formal notion of relevant information, it was physical reasoning that led us to an appropriate RG condition. By explicitly thinking about the properties physical systems possess, such as locality and translational invariance, we managed to peek deeper into 'black box' of neural networks than may have otherwise been possible. We quote Koch-Janusz and Ringel, "the internal data representations discovered by suitably designed algorithms are not just technical means to an end, but instead are a clear reflection of the underlying structure of the physical system" [5]. Integrating physical and ML perspectives proves a promising basis for a better insight into these algorithms. Not to mention, these techniques are unsupervised, so they should translate readily to problems other than the Ising model.

Comparisons and integrations of ML and RG are at an early phase, and there remains much to be uncovered. It seems that information theory is the appropriate framework with which to lay these links. Within physics, a better understanding of this area may guide research into new systems (on disordered, glassy systems and quantum field theories), though the ramifications extend much further [5]. Ultimately, the synthesis of ML, physics, and information theory may teach us a thing or two about why these techniques work as well as they do. Until then, the links should inspire powerful new techniques in both of these disciplines.

# Appendix A

## Basics of Statistical Physics

The proceeding derivations follow Cardy [12] and Domb [11].

### A.1 Measurements as Averages

Suppose we want to test whether our theory, the Ising model, lines up with experimental results: what are the kinds of predictions we can make? Our first attempt might be to measure the system's magnetization, rather appropriate as this is the defining characteristic of 'magnets.' Before we do, however, let us be more precise. What is a measurement, specifically for magnetization? The magnetization, $M$, of any one microstate, $\boldsymbol{s}$ is simply the sum of the orientations of all spins, $s_i \in \boldsymbol{s}$ in a given state:

$$M(\boldsymbol{s}) = \sum_i s_i, \tag{A.1}$$

Generally, the microscopic state of a system is changing rapidly, and the system will explore many different microstates over human time-scales. Our measuring devices have finite resolution in time, so, in a laboratory, we cannot perform sums over individual microstates. A measurement becomes the average over the microstates visited during the measurement period, weighted by the time spent in each state. In statistical physics, we often work in the limit that the update time goes to zero compared to the time of measurement. In this light, the notion of a probability of a microstate is also the fraction of time the system spends in that microstate over an infinite duration. Our measurement outcome, $\mathcal{M}$, is an expectation value, $\langle M(\boldsymbol{s})\rangle$, the average over the set of all microstates, $\mathcal{S} = \{\boldsymbol{s}\}$, weighted by the amount of time spent in each state, $P(s)$:

$$\mathcal{M} = \langle M(\boldsymbol{s})\rangle = \sum_{\boldsymbol{s}\in\{\boldsymbol{s}\}} P(\boldsymbol{s})M(\boldsymbol{s}).$$

(A.2)

The same holds for any measurable quantity. Given some function of a microstate, $A(\boldsymbol{s})$, the macroscopic value it corresponds to is its expectation value over all microstates, $\langle A(\boldsymbol{s})\rangle$. Common examples (not restricted to the Ising model) include $E(\boldsymbol{s})$, the energy, $P(\boldsymbol{s})$, the pressure, and $V(\boldsymbol{s})$, the volume. If we can find a probability measure over microstates, we can use the above to calculate any macroparameter we choose.

### A.2 Thermal Equilibrium and the Second Law of Thermodynamics

Consider the first law of thermodynamics (differential form with only one kind of particle):

$$dE = TdS - PdV + \mu dN,$$

(A.3)

where $E$ is the energy, $T$ the temperature, $P$ the pressure, $V$ the volume, $\mu$ the chemical potential, and $N$ the number of particles.

For the canonical ensemble (our example), $dV$ and $dN$ are $0$. Then, this reduces to:

$$\frac{1}{T}dE = dS.$$

(A.4)

From equation (2.6), we have $\Delta S = S_{\boldsymbol{r}}(\boldsymbol{s}) - S_{\boldsymbol{r}}(\boldsymbol{s}')$. If the reservoir is much larger than $\boldsymbol{s}$, this is a tiny difference, so $\Delta S \approx dS$. Then, we get $\Delta S \approx \frac{1}{T}(dE)$. Note that this requires the reservoir and system to have the same temperature (i.e., to be in thermal equilibrium, $T_{\boldsymbol{r}} = T_{\boldsymbol{s}}$). We can use the same approximation to expand this out: $dE \approx E_{\boldsymbol{r}}(\boldsymbol{s}) - E_{\boldsymbol{r}}(\boldsymbol{s}') = \Delta E$. We end up with equation (2.7).

Note that in the thermodynamic limit, $|\boldsymbol{s}|, |br| \to \infty$ while maintaining the inequality $1 \ll |\boldsymbol{s}| \ll |\boldsymbol{r}|$, these statements are precise (this is simply the central limit theorem).

### A.3 Observables as Derivatives of the Partition Function

Let us define the free energy, $F = -\frac{1}{\beta}\ln Z$, where $\beta = kT$, the so-called thermodynamic beta. Consider an energy function with a dependence on parameter $A$ according to:

$$E(\boldsymbol{s}) = E_0(\boldsymbol{s}) + \lambda A(\boldsymbol{s})$$

(A.5)

Then, $\langle A\rangle = \partial_\lambda F$. The derivation proceeds as:

$$\begin{aligned}
\langle A(\boldsymbol{s})\rangle &= \partial_\lambda\left(-\frac{1}{\beta}\ln Z\right)\\
&= -\frac{1}{\beta Z}\partial_\lambda Z\\
&= -\frac{1}{\beta Z}\sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}\partial_\lambda\left(-\beta E(\boldsymbol{s})\right)\\
&= \frac{1}{\beta Z}\sum_{\boldsymbol{s}} e^{-\beta E(\boldsymbol{s})}\left(\beta A(\boldsymbol{s})\right)\\
&= \sum_{\boldsymbol{s}} \frac{e^{-\beta E(\boldsymbol{s})}}{Z}A(\boldsymbol{s})\\
&= \sum_{\boldsymbol{s}} P(\boldsymbol{s})A(\boldsymbol{s}).
\end{aligned}$$

(A.6, A.7, A.8, A.9, A.10 & A.11, respectively)

Consider an example, taking the Ising Hamiltonian (2.10):

$$
\begin{aligned}
\partial_B\left(-\log Z\right) &= -\frac{1}{Z}\sum_{\boldsymbol{s}}\partial_B e^{-B\sum_i s_i - J\sum_{\langle i,j\rangle} s_i s_j} \\
&= -\sum_{\boldsymbol{s}}\frac{e^{-E(\boldsymbol{s})}}{Z}(-\sum_i s_i) \\
&= \sum_{\boldsymbol{s}} P(\boldsymbol{s})M(\boldsymbol{s}) \\
&= \langle M(\boldsymbol{s})\rangle.
\end{aligned}
$$

(A.12, A.13, A.14 & A.15, respectively)

Two others:

$$
\begin{aligned}
&\chi = \partial_B \mathcal{M} = \partial_B^2 F \\
&E = \partial_\beta F.
\end{aligned}
$$

(A.16 & A.17, respectively)

# Appendix B

## Scaling and Renormalization

### B.1 Finite-Size Scaling Analysis

This section follows [29]. In our theoretical discussion, we have typically considered infinite lattices: for a fixed lattice size $L$, we have let the microscopic size $a$ (the lattice width) go to zero. Then, the dimensionless $N = L/a$ (the number of lattice sites in a given dimension) diverges; this is the classical thermodynamic limit. Of course, truly physical systems are finite with correspondingly non-singular partition functions. For any such systems, there can be no divergences. An important concern is, then, determining under which conditions we can treat our systems as infinite and under which conditions finite-size effects have a non-negligible effect.

An important tool we use to study critical systems is MCMC simulations: here, we quickly run into computational limits on the sizes of lattices we can simulate. Especially for MCMC techniques, the role of finite-size scaling is crucial.

Consider a $d$-dimensional Ising model infinite in all directions. If we were to decrease $N$ of one dimension to some finite value, the system's critical behavior begins to be dominated by a $d-1$ dimensional critical point. This is an example of crossing-over behavior. By controlling some relevant parameter we change the effective dimensionality of our system and navigate between different critical points.

An example of this finite-size behavior is that of the correlation length, $\xi$. It will no longer diverge and depending on the boundary conditions will experience a peak at either above or below the infinite critical point. For cyclical boundary conditions, the effective critical temperature will increase, as the system is more ordered with more paths between spins. For zero-field boundary conditions, the temperature will decrease as there less paths between spins.

We have seen from Table 2.3, that diverging quantities $A(t, N^{-1}=0)$ scale as $|t-t_c|^{-\zeta}$ for some critical exponent $\zeta$. In the large $N$-limit, we expect that the system will continue to behave as such, provided that $N$ is much greater than the system's characteristic length scale, the correlation length, $\xi \sim |t-t_c|^{-\nu}$. This amounts to:

$$
A(t, N^{-1}) \sim |t-t_c|^{-\zeta} \sim \xi^{\zeta/\nu} \qquad (N \gg \xi, t \to t_c). \tag{B.1}
$$

The system's geometry will act as a cutoff: rather than diverge, now, as $\xi \to N$, $A$ will behave as

$$
A(t, N^{-1}) \sim N^{\zeta/\nu} \qquad (N \ll \xi, t \to t_c). \tag{B.2}
$$

Together, these considerations give rise to the finite-scaling ansatz:

$$
A(t, N^{-1}) \sim \xi^{\zeta/\nu}\phi((N\xi)^{-1}) \qquad (N^{-1} \to 0, t \to t_c), \tag{B.3}
$$

where

$$
\phi(x)\begin{cases} = \text{const.} & \text{for } |x| \gg 1 \\ \sim x^{\zeta/\nu} & \text{for } |x| \to 0 \end{cases}. \tag{B.4}
$$

$phi(x)$ is a scaling function that controls the finite-size effects. By convention, we choose the scaling function $\tilde{\phi}(x) = x^{-\zeta}\phi(x^{\nu})$. Then, we get that

$$A(t,L) = L^{\zeta/\nu}\tilde{\phi}\left(L^{1/\nu}(t-t_c)\right), \qquad (L \to \infty, t \to t_c), \tag{B.5}$$

with

$$\tilde{\phi}(x) \begin{cases} = \text{const.} & \text{for } x \to 0 \qquad (L \ll \xi), \\ \sim L^{-\zeta/\nu}(\varrho - \varrho_c)^{-\zeta} & \text{for } |x| \gg 1 \qquad (L \gg \xi). \end{cases}$$

.

(B.6)

This is valuable because it allows us to plot experimental data $a_{exp.}(t,L)$ collected at some temperature and lattice size, on a single curve:

$$\phi(L^{1/\nu}\tilde{(}t-t_c)) = L^{-\zeta/\nu}A(t,L)$$

.

(B.7)

Plotting Plotting $L^{1/\nu}(t - t_c)$ against $L^{-\zeta/\nu}A(t,L)|_{a_{exp.}(t,L)}$, our experimental data will 'collapse' onto a single line. In practice, it is enough to consider only the effective critical temperature at different length scales. Plotting this against $\log L$, the points should fall onto a single line with slope $\nu$.

### B.2 Scaling Rules for the Free-Energy

In this section, we will derive a transformation rule for the free energy. From Appendix A.3, we know that we can use this transformation rule to determine our critical exponents of interest.

Starting with equation (2.18), we can intuit that the free energy will take a form similar to:

$$f(\{K\}) = g(\{K\}) + b^{-d}f(\{K'\}). \tag{B.8}$$

we expect a function similar to our starting point but of the updated coupling $f(\{K'\})$ and where we have rescaled the system by (for example, a block-size) $b$. For $d$ dimensions, we rescale in each direction. Furthermore, we could have some (analytic) function of our coupling at a given instance, $g(\{K\})$.

The free energy of new system is equal to that of the original system with some rescaling, plus the addition of a constant term. Note that this is an inhomogeneous transformation. However, for *singular* behavior (near the critical point), we only care about the second term. $g()$ originates from summing over short degrees of freedom, and it should be an analytic (non-divergent) function of $K_a$ even at critical point.

Considering just the transformation rule for the singular part:

$$f_s(\{K\}) = b^{-d}f_s(\{K'\}). \tag{B.9}$$

We use our knowledge that there are only two relevant scaling variables, $u_t$ and $u_h$ (for the interested reader, we refer to Cardy [12]). These possess important symmetries: $u_t$ corresponds to the even subspace of couplings (invariant a collective sign-flip $s \to -s$), and $u_h$ to the odd subspace (equivariant with the collective sign-flip). Let us re-express the above equation in terms of our scaling variables. First, we Taylor-expand the scaling variables[14]. By these symmetry arguments, odd terms vanish from $u_t$ and even terms from $u_h$. Looking at the formula for $u_i$, we see they must vanish at $t = h = 0$.

$$u_t = \frac{t}{t_0} + O(t^2, h^2) \tag{B.10}$$

$$u_h = \frac{h}{h_0} + O(th) \tag{B.11}$$

Near the critical point, we take these scaling variables to be proportional to $h$ and $t$[15]. Combining our equations for the scaling variables, (B.10) and (B.11), with the singular free energy transformation:

$$f_s(u_t, u_h) = b^{-d}f_s(b^{y_t}u_t, b^{y_h}u_h). \tag{B.12}$$

If we repeat this transformation $n$ times, we get:

$$f_s(u_t, u_h) = b^{-nd}f_s(b^{ny_t}u_t, b^{y_h}u_h). \tag{B.13}$$

[14] This is valid since we have already limited ourselves to the immediate vicinity of the critical point.

[15] In more complicated systems, and those with tricritical points, our relevant variables will not necessarily be directly proportional to the experimental variables (the knobs) we control.

We now choose an arbitrary point $u_{t0} = |b^{ny_t} u_t|$. This constrains our $n$ to not be too large that the linear approximation breaks down. Solving for $n$ we get:

$$n = \frac{1}{y_t} \log_b \left( \left| \frac{u_t}{u_{t0}} \right|^{-1} \right). \tag{B.14}$$

Plugging equation (B.14) into equation (B.13):

$$f_s(u_t, u_h) = \left| \frac{u_t}{u_{t0}} \right|^{\frac{d}{y_t}} f_s(\pm u_{t0}, u_h \left| \frac{u_t}{u_{t0}} \right|^{-\frac{y_h}{y_t}}). \tag{B.15}$$

Rewriting in terms of $t$ and $h$, where we incorporate $u_{t0}$ into a new scale factor $t_0$ (i.e., $\frac{t}{t_0} \leftarrow \frac{t}{u_{t0} t_0}$):

$$f_s(t, h) = \left| \frac{t}{t_0} \right|^{d/y_t} \Phi(\frac{\frac{h}{h_0}}{|t/t_0|^{y_h/y_t}}) \tag{B.16}$$

$u_{t0}$ has allowed us to express the (approximate) function of two variables, $f_s$, in terms of just one variable in the *scaling function* $\Phi$.

This scaling function may appear to include a $u_{t0}$ dependency, but since the l.h.s. does not, this will cancel. These scaling functions turn out to be universal, it only depends on the system through $t_0$ and $h_0$.

### B.3 The Correlation Length Critical Exponent

As an example of using free energy scaling to solve critical exponents, let us attempt correlation length critical exponent. Consider the two-point correlation function:

$$G(r_1 - r_2, H) \equiv \langle s(r_1) s(r_2 \rangle_H - \langle s(r_1) \rangle_H \langle s(r_2) \rangle_H \tag{B.17}$$

Our first step will be to express this in terms of (derivatives of) the free energy. We introduce a non-uniform magnetic field to our Hamiltonian:

$$H \to H - \sum_r h(r) s(r) \tag{B.18}$$

and differentiate the free energy ($f = \ln Z$) twice:

$$\begin{aligned}
G(r_1 - r_2, H) =& \frac{\partial^2}{\partial h(r_1) \partial h(r_2)} \ln Z(\{H\})|_{h(r)=0} \\
=& \frac{\partial}{\partial h(r_1)} \left( -\frac{1}{Z} \sum_s s(r_2) e^{H(s) - \sum_r h(r) s(r)} \right) |_{h=0} \\
=& - \frac{1}{Z^2} \sum_{s'} s'(r_1) e^{H(s') - \sum_r h(r) s'(r)} \sum_s s(r_2) e^{H(s} \\
& - \frac{1}{Z} \sum_s s(r_1) s(r_2) e^{H(s) - \sum_r h(r) s(r)} |_{h=0} \\
=& - \left( \frac{1}{Z} \sum_{s'} s'(r_1) e^{H(s')} \right) \left( \frac{1}{Z} \sum_s s(r_1) e^{H(s)} \right) + \\
=& \langle s(r_1) s(r_2) \rangle_H - \langle s(r_1) \rangle_H \langle s(r_2) \rangle_H .
\end{aligned}$$

(B.19, B.20, B.21, B.22, B.23 & B.24, respectively)

We suppose that $h(r)$ varies over scales much larger than block size $ba$. Applying block renormalization we can effectively ignore the varying of $h(r)$ within a given block. Then, for a specified block, it transforms as a uniform field would. The renormalization Hamiltonian should be of the same form.

$$H'(s') - \sum_{r'} h'(r') s'(r') \tag{B.25}$$

Since RG preserves the partition function, we can write:

$$\frac{\partial^2 \ln Z'(h')}{\partial h'(r_1') \partial h'(r_2')} = \frac{\partial^2 \ln Z(h)}{\partial h'(r_1') \partial h'(r_2')} \tag{B.26}$$

Now the l.h.s. is just the correlation function of the RG system with the new Hamiltonian. In terms of our original correlation function, we have to rescale the distance by a factor b.

$$G((r_1 - r_2)/b, H') \tag{B.27}$$

Onto the r.h.s. If we make a change in $h'(r')$, we will change *all* the spins contained in that block:

$$\delta h(r_i) = b^{-y_h} \delta h'(r_1'), \tag{B.28}$$

so this reduces to:

$$b^{-2y_h}\langle(s_1^{(1)}+s_2^{(1)}+\dots)(s_1^{(2)}+s_2^{(2)}+\dots)\rangle_H. \tag{B.29}$$

Each block contains $b^d$ spins, so expanding this gives us $b^{2d}$ 2-point correlations. For our assumption that $r =|r_1-r_2|$ is large, each will give about the same result. For isotropic systems (respecting rotational symmetries of the lattice), the correlation function will only depend on distance between points, not on orientation.

We end up with the transformation rule:

$$G((r_1-r_2)/b, H') = b^{2(d-y_h)}G(r_1-r_2, H). \tag{B.30}$$

For simplicity, we set $h=0$ and iterate $n$ times:

$$G(r,t) = b^{-2(d-y_h)}G(r/b, b^{y_t}t) = b^{-2n(d-y_h)}G(r/b^n, b^{ny_t}t), \tag{B.31}$$

stopping at some fixed point where $b^{ny_t}(t/t_0)=1$ (as we did for the free energy in Appendix B.2). Then,

$$n = -\frac{1}{y_t}\log_b(t/t_0), \tag{B.32}$$

and plugging this in:

$$G(r,t) = |t/t_0|^{2(d-y_h)/y_t}\Psi\left(\frac{r}{|t/t_0|^{-1/y_t}}\right). \tag{B.33}$$

By definition, $G(r) \propto e^{-r/\xi}$ for $r \gg \xi$, near the critical point (see argument in, for example, [12]). From the argument of the scaling function, $Psi$, we can identify that: $\xi \propto |t|^{-1/y_t}$.

Then,

$$\nu = 1/y_t \tag{B.34}$$

# Appendix C

## Basics of Information Theory

We begin with a random variable $x \in \mathcal{X}$, with the probability distribution $P(x)$. The *information* in $x$ is:

$$I(x) = -\log_2 P(x). \tag{C.1}$$

To understand why this is more useful than $P(x)$, consider another random variable $y \in \mathcal{Y}$ with distribution $P(y)$. If the two variables are independent ($P(x,y) = P(x)P(y)$), we get that their information *adds*:

$$I(x,y) = I(x) + I(y). \tag{C.2}$$

The logarithm is a powerful tool that converts multiplication (hard) to addition (easy). Intuitively, it makes sense. An event which is $100\%$ likely $P(x)=1$, is guaranteed to happen, so when it happens, it carries no information. An event which is infinitely unlikely, $P(x)=0$ should never happen. If it does, somehow, that event carries an infinite amount of information. Precisely, information is measured in bits a $0$ or $1$. Consider trying to encode samples of $P(x)$ as effectively as possible. When $P(x)=1/2$, we need only one bit to encode $x$. For $P(x)=1/4$, two bits, and so on.

In this light, entropy is the expected value of information: were we to sample $P(x)$, then what is the average information we glean from any one sample? In this limiting case of infinite samples, we perform a sum of the information of each state weighted by the probability of that state:

$$S(\mathcal{X}) = \langle I(x)\rangle = \sum_{x\in\mathcal{X}} P(x)I(x). \tag{C.3}$$

### C.1 Cross-Entropy

Assume we have one set of events, $\mathcal{X}$, but two distributions $P(x)$ and $Q(x)$. We have devised an optimal encoding of the events in $\mathcal{X}$ using $Q(x)$. This means we need the least amount of bits that is possible to identify events $x$ from $Q(x)$. If instead, the events suddenly come from $P(x)$ and $P(x) \neq Q(x)$, we will need, on average, more bits to identify $x$.

The *average* number of bits we need is the *cross-entropy*, $H$, identified as the expectation over $P$ of the information over $Q$:

$$H(P,Q) = \langle I_Q(x)\rangle_P = \sum_{x\in\mathcal{X}} P(x)I_Q(x) = -\sum_x P(x)\log Q(x).$$

(C.4)

Choosing $Q$ so as to minimize this quantity, we derive an encoding closer to $P(x)$.

### C.2 Kullback-Leibler Divergence

If we rewrite the KLD (3.13) as

$$D_{KL}(P||Q) = H(P,Q) - S(P),$$

(C.5)

$P$ and $Q$ minus the entropy over $P$. In other words this is the average number of *additional* bits we need to identify $x$ from $P$ when using the improper coding $Q$. This amount is $0$ if and only if $H(P,Q) = S(P)$. From equation (C.4) and equation (C.3), we see, by the positive semi-definiteness of entropy, this immediately implies that $P$=$Q$.

# Appendix D

## Restricted Boltzmann Machines

### D.1 Factoring of the Marginal Distribution

Our aim is to solve for $E_\theta(\boldsymbol{v})$, the energy function describing the marginalized system of visibles of an RBM, equation (3.12), in terms of the joint energy function of the full system, equation (2.10). Combining these two equations, we get the following.

$$\begin{aligned} e^{-E_\theta(\boldsymbol{v})} &= \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{v},\boldsymbol{h})} \\ \iff E_\theta(\boldsymbol{v}) &= -\log\left(\sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{v},\boldsymbol{h})}\right), \end{aligned}$$

(D.1 & D.2, respectively)

where the partition functions have cancelled out. We can rewrite the right hand side:

$$\begin{aligned} \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{v},\boldsymbol{h})} &= \sum_{\boldsymbol{h}} e^{-\sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} v_i w_{ij} h_j} \\ &= e^{-\sum_i a_i v_i} \sum_{\boldsymbol{h}} e^{-\sum_j \left(b_j + \sum_i v_i w_{ij}\right) h_j}. \end{aligned}$$

(D.3, D.4 & D.5)

Furthermore,

$$\begin{aligned} \sum_{\boldsymbol{h}} e^{-\sum_j \left(b_j+\sum_i v_i w_{ij}\right)h_j} &= \sum_{\boldsymbol{h}} \prod_j e^{-\left(b_j+\sum_i v_i w_{ij}\right)h_j} \\ &= \prod_j \sum_{h_j\in\{0,1\}} e^{-\left(b_j+\sum_i v_i w_{ij}\right)h_j} \\ &= \prod_j \left(1 + e^{-\left(b_j+\sum_i v_i w_{ij}\right)}\right). \end{aligned}$$

(D.6, D.7 & D.8, respectively)

Plugging in, we see,

$$\begin{aligned} E_\theta(\boldsymbol{v}) &= -\log\left(e^{-\sum_i a_i v_i} \prod_j \left(1 + e^{-\left(b_j+\sum_i v_i w_{ij}\right)}\right)\right) \\ &= \sum_i a_i v_i - \sum_j \log\left(1 + e^{-\left(b_j+\sum_i v_i w_{ij}\right)}\right). \end{aligned}$$

(D.9 & D.10, respectively)

### D.2 Correspondence between Kadanoff's Variational RG and RBMs

When we form a comparison with restricted Boltzmann machines, we will consider the limited case of exact transformations (i.e., $\Delta F = 0 \iff F(\boldsymbol{v}) = F_\theta(\boldsymbol{h})$). This implies that:

$$\begin{aligned} F(\boldsymbol{v}) &= F_\lambda(\boldsymbol{h}) \\ -\ln\left(\sum_{\boldsymbol{v}} e^{-\boldsymbol{H}(\boldsymbol{v})}\right) &= -\ln\left(\mathrm{Tr}_{h_j,v_i}\, e^{\boldsymbol{T}_\lambda(\boldsymbol{v},\boldsymbol{h})-\boldsymbol{H}(\boldsymbol{v})}\right) \\ \sum_{\boldsymbol{v}} e^{-\boldsymbol{H}(\boldsymbol{v})} &= \mathrm{Tr}_{h_j,v_i}\, e^{\boldsymbol{T}_\lambda(\boldsymbol{v},\boldsymbol{h})-\boldsymbol{H}(\boldsymbol{v})} \\ &= \sum_{\boldsymbol{v}} \sum_{\boldsymbol{h}\in\mathcal{H}} e^{\boldsymbol{T}_\lambda(,\boldsymbol{h})} e^{-\boldsymbol{H}(\boldsymbol{v})} \\ &= \sum_{\boldsymbol{v}} e^{-\boldsymbol{H}(\boldsymbol{v})} \sum_{\boldsymbol{h}\in\mathcal{H}} e^{\boldsymbol{T}_\lambda(\boldsymbol{v},\boldsymbol{h})}, \end{aligned}$$

(D.11, D.12, D.13, D.14 & D.15, respectively)

which is true if and only if:

$$\mathrm{Tr}_{h_j}\, e^{\boldsymbol{T}_\lambda(\boldsymbol{v},\boldsymbol{h})} = 1.$$

(D.16)

We know that $P^{(RBM)}(\boldsymbol{h})$ is a marginalization over the energy function $E(\boldsymbol{v},\boldsymbol{h})$. Combining this with equation (D.6):

$$\begin{aligned} P(\boldsymbol{h}) = \frac{e^{-\boldsymbol{H}^{RBM}(\boldsymbol{h})}}{Z^h} &= \sum_{v_i} \frac{e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})}}{Z} \\ &= \sum_{v_i} \frac{e^{\boldsymbol{T}(\boldsymbol{v},\boldsymbol{h})-\boldsymbol{H}(\boldsymbol{v})}}{Z} \\ &= \frac{e^{-\boldsymbol{H}^{RG}(\boldsymbol{h})}}{Z^h}. \end{aligned}$$

(D.17, D.18 & D.19, respectively)

We identify:

$$\boldsymbol{H}^{RBM}(\boldsymbol{h}) = \boldsymbol{H}^{RG}(\boldsymbol{h})$$

(D.20)

The Hamiltonian describing our hidden spins after block-spin renormalization also describes the configuration of hidden spins in our RBM. Furthermore, we can expand:

$$\begin{aligned} e^{\boldsymbol{T}(\boldsymbol{v},\boldsymbol{h})} &= e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})+\boldsymbol{H}(\boldsymbol{v})} \\ &= \frac{e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})}}{\sum_{v_i,h_j} e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})}} \cdot \left(\sum_{v_i,h_j} e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})}\right) \cdot \frac{P(\boldsymbol{v})}{P(\boldsymbol{v})} \cdot e^{-\boldsymbol{H}(\boldsymbol{v})} \\ &= \frac{P(\boldsymbol{v},\boldsymbol{h})}{P(\boldsymbol{v})} \cdot \frac{\sum_{v_i,h_j} e^{-\boldsymbol{E}(\boldsymbol{v},\boldsymbol{h})}}{\sum_{v_i} e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})}} e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})-\boldsymbol{H}(\boldsymbol{v})} \\ &= P(\boldsymbol{h}|\boldsymbol{v}) \cdot \frac{\sum_{v_i} e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})}}{\sum_{v_i} e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})}} e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})-\boldsymbol{H}(\boldsymbol{v})} \\ &= P(\boldsymbol{h}|\boldsymbol{v}) \cdot e^{-\boldsymbol{H}^{RBM}(\boldsymbol{v})-\boldsymbol{H}(\boldsymbol{v})}. \end{aligned}$$

(D.21, D.22, D.23, D.24 ˆ D.25)

This lends Kadanoff's visible-hidden coupling $\boldsymbol{T}$ an interpretation as a kind of variational conditional probability distribution. Using the exact case, equation (D.16), we show:

$$\begin{aligned} 1 &= \sum_{h_j} e^{\boldsymbol{T}_\lambda(\boldsymbol{v},\boldsymbol{h})} \\ &= \sum_{h_j} P(\boldsymbol{v}|\boldsymbol{h}) \frac{P(\boldsymbol{v})}{P_{true}(\boldsymbol{v})} \\ &= \frac{P(\boldsymbol{v})}{P_{true}(\boldsymbol{v})}. \end{aligned}$$

(D.26, D.27 & D.28, respectively)

.

Therefore, $P(\boldsymbol{v}) = P_{true}(\boldsymbol{v})$. This derivation trivially goes both ways.

### D.3 An Exact Correspondence between Majority-Rule RG and Convolutional RBMs

In this section, we will see that it is possible to implement a majority-rule block update with an RBM. Note that this does not concern whether or not any given RBM will actually learn this rule. For arguments considered in Chapter 4, we need only consider the update function for a single block, $P(h_j|\boldsymbol{v}^j)$. Ignoring the index on $\boldsymbol{v}^{(j)}$, let us remind ourselves of the RBM's conditional distribution:

$$P(h_j|\boldsymbol{v}) = \frac{1}{1+e^{-h_j(\sum_i w_{ij}v_i+b_j)}}.$$

(D.29)

For block renormalization, we want each spin $v_i$ to contribute equally, so we can consider the simpler $w_{ij} \to w_j$. Let,

$$-h_j(w_j \sum_i v_i + b_j).$$

(D.30)

If we have $N$, visible spins (i.e. $i \in \{1 \dots N\}$), then we can rewrite the sum over visibles in terms of their average $\langle v_i \rangle$ as

$$-h_j w_j \left( N\langle v_i \rangle + b_j \right).$$

(D.31)

If we choose $b_j = -Nw_j/2$, we get:

$$-h_j w_j N \left( \langle v_i \rangle - \frac{1}{2} \right).$$

(D.32)

Restricting our attention to $\langle v_i \rangle - \frac{1}{2}$, we see this is positive if and only if more than half of the spins in $v_i$ are up, $1$, and negative if and only if half the spins are down, $0$. Narrowing our attention to $\langle h_j \rangle = P(h_j|\boldsymbol{v})$, the trick to recover the majority rule transformation will be to let $w_j \to -\infty$.

$$\langle h_j \rangle = \lim_{w_j \to \infty} \frac{1}{1 + e^{-w_j N\left(\langle v_i \rangle - \frac{1}{2}\right)}}$$

(D.33)

We distinguish three cases:

$$\langle h_j \rangle = \begin{cases} 0 & \iff \langle v_i \rangle < 1/2 \\ 0.5 & \iff \langle v_i \rangle = 1/2 \\ 1 & \iff \langle v_i \rangle > 1/2. \end{cases}$$

(D.34)

This is exactly the majority-rule, including even the probabilistic update rule for blocks of even numbers of spins.

# Appendix E

## The Real-Space Mutual Information Maximization Algorithm

### E.1 A Proxy for the Mutual Information

As a first step, let us factor the joint distribution in equation (5.1) ($\boldsymbol{h}$ on $\boldsymbol{e}$ is mediated entirely through $\boldsymbol{v}$):

$$P(\boldsymbol{h}, \boldsymbol{e}) = \sum_{\boldsymbol{v}} P(\boldsymbol{h}|\boldsymbol{v}) P(\boldsymbol{v}, \boldsymbol{e}).$$

(E.1)

In Chapter 6, we saw that we can interpret an RG transformation as a conditional probability distribution $P(\boldsymbol{h}|\boldsymbol{v})$, equation (4.4) (dropping the index $j$ and allowing multiple hidden units per block). As we discussed in Chapter 5, we model this distribution with an RBM, with parameters $\Lambda$, such that:

$$P_\Lambda(\boldsymbol{h}, \boldsymbol{v}) = \frac{e^{-E_\Lambda(\boldsymbol{h}, \boldsymbol{v})}}{\sum_{\boldsymbol{h}', \boldsymbol{v}'} e^{-E_\Lambda(\boldsymbol{h}', \boldsymbol{v}')}},$$

$$E_\Lambda(\boldsymbol{h}, \boldsymbol{v}) = -\left( \sum_{ij} w_{ij}^{(\Lambda)} v_i h_j + \sum_i a_i^{(\Lambda)} v_i + \sum_j b_j^{(\Lambda)} h_j \right).$$

(E.2 & E.3, respectively)

As we saw, we can use the above to derive easy to evaluate equations for $P_\Lambda(\boldsymbol{h}|\boldsymbol{v})$, $P_\Lambda(\boldsymbol{h})$, and $P_\Lambda(\boldsymbol{v})$. In fact, we can also derive an equation for $P_\Lambda(\boldsymbol{v}|\boldsymbol{h})$. However, for RG, we only care about transformations in one direction, $\boldsymbol{v} \to \boldsymbol{h}$. Therefore, for simplicity, we can set $a_i^{(\Lambda)} = 0$ (this term factors out in $P_\Lambda(\boldsymbol{h}|\boldsymbol{v})$). We get that:

$$P_\Lambda(\boldsymbol{v}) = \sum_{\boldsymbol{h} P_\Lambda(\boldsymbol{h}, \boldsymbol{v})} = \frac{e^{-E_\Lambda(\boldsymbol{v})}}{\sum_{\boldsymbol{v} e^{-E_\Lambda(\boldsymbol{v})}}}$$

$$E_\Lambda(\boldsymbol{v}) = -\sum_j \log\left(1 + e^{-\left(b_j + \sum_i v_i w_{ij}\right)}\right),$$

(E.4 & E.5, respectively)

and

$$P_\Lambda(\boldsymbol{h}|\boldsymbol{v}) = \frac{P_\Lambda(\boldsymbol{h}, \boldsymbol{v})}{P_\Lambda(\boldsymbol{v})} = e^{-E_\Lambda(\boldsymbol{h}, \boldsymbol{v}) + E_\Lambda(\boldsymbol{v})}.$$

(E.6)

Combining with our mutual information condition, Equation 5.1, we get:

$$A_\Lambda(\boldsymbol{h} : \boldsymbol{e}) = \sum_{\boldsymbol{h}, \boldsymbol{v}, \boldsymbol{e}} P_\Lambda(\boldsymbol{h}|\boldsymbol{v}) P(\boldsymbol{v}, \boldsymbol{e}) \log \left( \frac{\sum_{\boldsymbol{v}} P(\boldsymbol{v}, \boldsymbol{e}) P_\Lambda(\boldsymbol{h}|\boldsymbol{v})}{\sum_{\boldsymbol{v}', \boldsymbol{e}} P(\boldsymbol{v}', \boldsymbol{e}) P_\Lambda(\boldsymbol{h}|\boldsymbol{v}')} \right)$$

(E.7)

We assume that all of these distributions (not only those defined by the $\Lambda$-RBM) are of Boltzmann form. Then, we can factor the partition functions, keeping an equation of Boltzmann form: Since these probabilities are all of Boltzmann form, we can factor out partition functions, leaving behind an equation with Boltzmann terms:

$$\frac{\sum_{\boldsymbol{v}} P(\boldsymbol{v},\boldsymbol{e})P_\Lambda(\boldsymbol{h}|\boldsymbol{v})}{\sum_{\boldsymbol{v}'} P(\boldsymbol{v}')P_\Lambda(\boldsymbol{h}|\boldsymbol{v}')} \to \frac{\sum_{\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{e})-E_\Lambda(\boldsymbol{h},\boldsymbol{v})+E_\Lambda(\boldsymbol{v})}}{\sum_{\boldsymbol{v}'} e^{-E(\boldsymbol{v}')-E_\Lambda(\boldsymbol{h},\boldsymbol{v}')+E_\Lambda(\boldsymbol{v}')}}.$$

(E.8)

We can reexpress the argument of the logarithm as follows:

$$\frac{\sum_{\boldsymbol{v}} e^{-E(\boldsymbol{v},\boldsymbol{e})-E_\Lambda(\boldsymbol{h},\boldsymbol{v})+E_\Lambda(\boldsymbol{v})}}{\sum_{\boldsymbol{v}'} e^{-E(\boldsymbol{v}')-E_\Lambda(\boldsymbol{h},\boldsymbol{v}')+E_\Lambda(\boldsymbol{v}')}} = \frac{\sum_{\boldsymbol{v}} e^{-E(\boldsymbol{v})-E_\Lambda(\boldsymbol{h},\boldsymbol{v})+E_\Lambda(\boldsymbol{v})} e^{-E(\boldsymbol{v},\boldsymbol{e})+E(\boldsymbol{v})}}{\sum_{\boldsymbol{v}'} e^{-E(\boldsymbol{v}')-E_\Lambda(\boldsymbol{h},\boldsymbol{v}')+E_\Lambda(\boldsymbol{v}')}}. = \frac{\sum_{\boldsymbol{v}} e^{-E_\Lambda(\boldsymbol{h},\boldsymbol{v},\boldsymbol{e})} e^{-\Delta E(\boldsymbol{v},\boldsymbol{e})}}{\sum_{\boldsymbol{v}'} e^{-E_\Lambda(\boldsymbol{h},\boldsymbol{v},\boldsymbol{e})}},$$

(E.1)

where

$$E_\Lambda(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h}) = E(\boldsymbol{v},\boldsymbol{e}) + E_\Lambda(\boldsymbol{h},\boldsymbol{v}) - E_\Lambda(\boldsymbol{v})$$
$$\Delta E(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h}) = E(\boldsymbol{v},\boldsymbol{e}) - E(\boldsymbol{v}).$$

(E.10 & E.11, respectively)

This is the expectation of $e^{-\Delta E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{e})}$ over the Boltzmann distribution with energy $E_\Lambda(\boldsymbol{h},\boldsymbol{v},\boldsymbol{e})$, where $\boldsymbol{h}$ and $\boldsymbol{e}$ are clamped. We write $\langle e^{-\Delta E(\boldsymbol{v},\boldsymbol{h},\boldsymbol{e})}\rangle_\Lambda[\boldsymbol{e},\boldsymbol{h}]$, where $[\boldsymbol{e},\boldsymbol{h}]$ denotes the dependence of this expectation value on the clamped values. Although $\Delta E$ has no dependence on $\Lambda$, its expectation value gains a dependence through $P_\Lambda(\boldsymbol{h},\boldsymbol{v},\boldsymbol{e})$. We get the following expression for the mutual information proxy:

$$A_\Lambda(\boldsymbol{h}:\boldsymbol{e}) = \sum_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{e}} P_\Lambda(\boldsymbol{h}|\boldsymbol{v})P(\boldsymbol{v},\boldsymbol{e}) \log\left(\langle e^{-\Delta E(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})}\rangle_\Lambda[\boldsymbol{e},\boldsymbol{h}]\right),$$

(E.12)

Here we see that the clamped values of $\boldsymbol{e}$ and $\boldsymbol{h}$ are given by the outside sum, so we write $[\boldsymbol{e},\boldsymbol{h}]$ after the expectation value to denote its dependence on these variables.

To further simplify this expression, we use a cumulant expansion:

$$\langle e^{K(\boldsymbol{x})}\rangle = e^{\sum_{\kappa=0}^{\infty}\frac{1}{\kappa!}C_\kappa},$$

(E.13)

with the cumulants expressed in terms of the moments. The first three terms are:

$$C_1 = \langle K\rangle$$
$$C_2 = \langle K^2\rangle - \langle K\rangle^2$$
$$C_3 = \langle K^3\rangle - 3\langle K^2\rangle\langle K\rangle + 2\langle K\rangle^3.$$

(E.14, E.15 & E.16, respectively)

This extends to distributions that include a dependence on other variables (i.e., $\boldsymbol{e}$, $\boldsymbol{h}$). Then, we can approximate:

$$A_\Lambda(\boldsymbol{h}:\boldsymbol{e}) \approx \sum_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{e}} P_\Lambda(\boldsymbol{h}|\boldsymbol{v})P(\boldsymbol{v},\boldsymbol{e})\langle -\Delta E(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})\rangle[\boldsymbol{e},\boldsymbol{h}],$$

(E.17)

where now

$$\langle \Delta E_\Lambda(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})\rangle[\boldsymbol{e},\boldsymbol{h}] \equiv \frac{\sum_{\boldsymbol{v}} \left(\Delta E_\Lambda(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})\right) e^{-E_\Lambda(\boldsymbol{v},\boldsymbol{h})}}{\sum_{\boldsymbol{v}'} e^{-E_\Lambda(\boldsymbol{v}',\boldsymbol{h})}}.$$

(E.18)

In general, we will not have access to the energy functions $E(\boldsymbol{v})$ and $E(\boldsymbol{v},\boldsymbol{e})$. Even if we were to have access to $P(\boldsymbol{x} = \{\boldsymbol{v},\boldsymbol{b},\boldsymbol{e},\boldsymbol{o}\})$, the necessary marginalizations ($P(\boldsymbol{v},\boldsymbol{e}) = \sum_{\boldsymbol{b},\boldsymbol{o}} P(\boldsymbol{x})$ and $P(\boldsymbol{v}) = \sum_{\boldsymbol{b},\boldsymbol{e},\boldsymbol{o}} P(\boldsymbol{x})$) are not necessarily tractable calculations. Therefore, we approximate $E(\boldsymbol{v},\boldsymbol{e})$ and $E(\boldsymbol{v})$ with two other RBMs with parameters $\Theta$ and $\Psi$, respectively:

$$E(\boldsymbol{v},\boldsymbol{e}) \approx E_\Theta(\boldsymbol{v},\boldsymbol{e})$$
$$E(\boldsymbol{v}) \approx E_\Psi(\boldsymbol{v}).$$

(E.19 & E.20, respectively)

Note that $(\boldsymbol{v},\boldsymbol{e})$ and $(\boldsymbol{v})$ are the inputs, the visible layers of these two RBMs. We introduce a second, hidden layer, which we marginalize over to get the above quantities.

Plugging in our learned distributions, we get:

$$A_{\Lambda,\Theta,\Psi}(\boldsymbol{h}:\boldsymbol{e}) \approx \sum_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{e}} P_\Lambda(\boldsymbol{h}|\boldsymbol{v}) P(\boldsymbol{v},\boldsymbol{e}) \langle -\Delta E_{\Theta,\Psi}(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})\rangle_{\Lambda,\Theta,\Psi}[\boldsymbol{e},\boldsymbol{h}], \tag{E.21}$$

We have derived an expression which we evaluate using Monte Carlo averages.

$$A_{\Lambda,\Theta,\Psi}(\boldsymbol{h}:\boldsymbol{e}) \approx \frac{1}{N^{(\boldsymbol{v},\boldsymbol{e})}} \sum_{\boldsymbol{h},\boldsymbol{v},\boldsymbol{e}} P_\Lambda(\boldsymbol{h}|\boldsymbol{v}) \langle -\Delta E_{\Theta,\Psi}(\boldsymbol{v},\boldsymbol{e},\boldsymbol{h})\rangle_{\Lambda,\Theta,\Psi}[\boldsymbol{e},\boldsymbol{h}] \tag{E.22}$$

First, we generate samples of $\boldsymbol{e}$ and $\boldsymbol{v}$ simply from partitions of our dataset $\boldsymbol{x}$[16]. Then, we use our $\Lambda$-RBM to translate samples of $\boldsymbol{v}$ into samples of $\boldsymbol{h}$. For each combination of $\boldsymbol{e}$ and $\boldsymbol{h}$, we generate wholly new samples, $\boldsymbol{v}'$, with the energy function $E_{\Lambda,\Psi}(\boldsymbol{v},\boldsymbol{e})$, over which we perform the internal MC-average.

In fact, we are not interested in the quantity, $A_{\Lambda,\Theta,\Psi}(\boldsymbol{h}:\boldsymbol{e})$ as much as we are in its derivative. With some simple (though exceedingly tedious)[17] Algebra, we can evaluate an expression for the derivative with respect to $\Lambda$of our mutual information proxy. It is crucial we calculate this explicitly, because through the stochastic nature of MC-sampling, our proxy $A_\Lambda$ will often have a zero-gradient. We would not be able to use stochastic gradient descent.

*Comment on the Cumulant Expansion. It is not entirely clear why the authors felt this expansion was necessary. The expectations are ultimately calculated over Monte Carlo samples, and $\langle K(\boldsymbol{x})\rangle$ is not much of an improvement over $\langle e^{K(\boldsymbol{x})}\rangle$ in computational complexity. Furthermore, it disregards information about higher order terms. It may be that this step manages to suppress irrelevant fluctuations, but a more rigorous understanding is needed. In the future, we may implement the algorithm, including additional terms in this expansion, comparing the ultimate performance against an un-expanded baseline. Then, we can more rigorously justify or critique this assumption.

### E.2 Intrinsic Thermometer

In order to measure critical exponents, we need a means of measuring the values of macrovalues through our iterations. If our aim is to measure temperature, we may, however, not have access to an explicit thermometer. To further complicate matters, as we saw previously, RG transformations generally introduce higher order correlations. From data alone, it is not necessarily possible to identify the contributions from different terms in the Hamiltonian. For our approach to be general, then, we need a means of implicit macroparameter calculations. With regard to temperature, we have several options.'

1. First they looked at using the MC configurations at given iterations. We can compute expectations of functions like the nearest-neighbor and next-nearest-neighbor correlation. Since these depend monotonically on the temperature near the critical point, we can use these to recreate an effective temperature at successive length scales.
2. Next, they used the mutual information as a proxy to the temperature. This, too, depends monotonically on the temperature near the critical point.
3. Finally, they mentioned the possibility of using the $\Lambda$- and $\Psi$-RBMs. These we can also use to measure expectations of correlations, and we can even intrinsically evaluate effective temperatures.

There are yet other options, not considered by Koch-Janusz and Ringel. We can formulate the task of measuring the effective temperature of a set of samples as a supervised learning problem. Then, we can further leverage the power of neural networks to act as our thermometers[18] So far, we have only considered neural networks which take a fixed number of inputs. Our aim for a NN thermometer would be that it could accurately measure the temperature of systems with different numbers of spins (at different RG steps). Two immediate solutions come to mind. First, we could train the networks on subsets of the $x$ samples. Then, we would train a separate network for each successive step. Second, we could

[16] We could have generated these samples de novo using the $\Theta$-RBM, but that would introduce needless time-complexity. Instead, we use that data we already have access to

[17] I mean whiteboards and whiteboards of it.

[18] Accomplished for example by Iso et al. [6]).

make use of recursive neural networks[19]. These recursive neural networks employ, as their name suggests, recursive weight-sharing, which explicitly allows for variable-sized inputs.

However, this would reduce the extent to which the RSMI algorithm is truly unsupervised. In practice, this need not be an issue: we assume that we have access to some data, and more often than not, this will be given to us by MCMC techniques. Then, we necessarily have a means of measuring parameters like these, at least in the non-coarse-grained systems.

Though we did not have the time to implement these ideas, we will continue developing *rgpy*, and we intend to introduce these features in later releases.

For our implementation, we used the next-nearest correlation function. For each value of this function we derive in successive iterations, we assign the temperature that it is closest too in our sample set. In the future, we will use more complicated functions: this ‘nearest-neighbors’ approach introduces significant error margins.

### E.3 Experimental Realization

Ultimately, our calculation of the correlation length critical exponent, $\nu$, proceeded as follows. We generated samples ($1000$ per temperature) of Ising configurations of various lattice widths ($8$, $16$, $32$, $64$). From values for the next-nearest neighbor correlation function, we devised a “thermometer”, see preceding section, for each length-scale. We trained the RSMI algorithm on samples of 64-by-64 spins for a total of 3 RG iterations. Each RBM we trained for 30 epochs. For the rest, our experimental set-up mirrored that of Koch-Janusz and Ringel **?**. Having generated results, and measurements of the magnetic susceptibility, $\chi$, for each step in each temperature sequence. From the peaks in susceptibility, we identified the effective critical temperature. Plotting this on one curve against $1/L$, using arguments from finite-size scaling (see section **??**), we collapsed the data onto a single curve whose slope equals $\nu$.

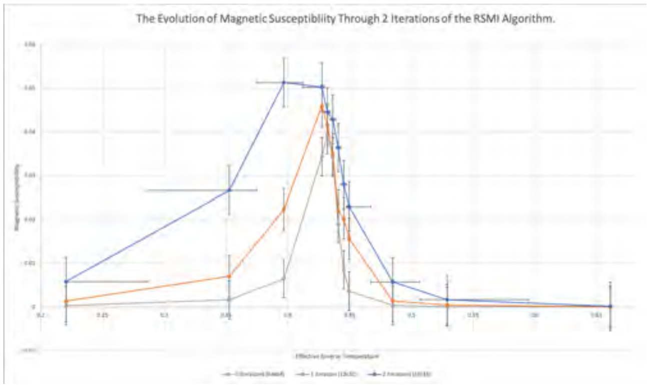


*Figure E.1 – The Magnetic Susceptiblity, $\chi$, at different iterations of the RSMI algorithm.*

## Works Cited

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[2] Henry W. Lin, Max Tegmark, and David Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 168:1223–1247, Sep 2017.

[3] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[4] Pankaj Mehta and David J. Schwab. An exact mapping between the variational renormalization group and deep learning. *arXiv e-prints*, page arXiv:1410.3831, Oct 2014.

[5] Maciej Koch-Janusz and Zohar Ringel. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14:578–582, Jun 2018.

[6] Satoshi Iso, Shotaro Shiba, and Sumito Yokoo. Scale-invariant feature extraction of neural network and renormalization group flow. *arXiv e-prints*, 97:053304, May 2018.

[7] David J. Schwab and Pankaj Mehta. Comment on ”why does deep and cheap learning work so well?” [arxiv:1608.08225]. *arXiv e-prints*, page arXiv:1609.03541, Sep 2016.

[8] Jaco ter Hoeve. Renormalization group connected to neural networks, 2018.

[9] Patrick M. Lenggenhager, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. Optimal renormalization group transformation from information theory. *arXiv e-prints*, page arXiv:1809.09632, Sep 2018.

[19] Not to be confused with recurrent neural networks, a particular kind of recursive neural network used for processing 1-dimensional (typically temporal) data.

[10] Jesse Hoogland. rgpy. https://github.com/jqhoogland/rgpy, 2019.

[11] Cyril Domb. *The critical point: a historical introduction to the modern theory of critical phenomena*. CRC Press, 1996.

[12] John Cardy. *Scaling and Renormaliztion in Statistical Physics*. Cambridge University Press, Cambridge, United Kingdom, 1996.

[13] JGTechSol. Optical computing, Feb 2017.

[14] Alan H Guth. Time since the beginning. *arXiv preprint astro-ph/0301199*, 2003.

[15] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science Business Media, 2013.

[16] Jian-Sheng Wang and Robert H Swendsen. Cluster monte carlo algorithms. *Physica A: Statistical Mechanics and its Applications*, 167(3):565–579, 1990.

[17] Kenneth G Wilson. Problems in physics with many scales of length. *Scientific American*, 241(2):158–179, 1979.

[18] Michael E Peskin. *An introduction to quantum field theory*. CRC Press, 2018.

[19] Leo P Kadanoff, Anthony Houghton, and Mehmet C Yalabik. Variational approximations for renormalization group transformations. *Journal of Statistical Physics*, 14(2):171–203, 1976.

[19] Geoffrey E Hinton. A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade*, pages 599–619. Springer, 2012.

[20] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *arXiv e-prints*, page arXiv:1803.08823, Mar 2018.

[21] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[22] Mukul Rathi. Learning through gradient descent, Aug 2018.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[24] Max Welling, Michal Rosen-Zvi, and Geoffrey E Hinton. Exponential family harmoniums with an application to information retrieval. In Advances in neural information processing systems, pages 1481–1488, 2005.

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[26] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[27] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

[28] Andreas Sorge. (2015). pyfssa 0.7.6. Zenodo. DOI: https://pyfssa.readthedocs.io/en/stable/fss-theory.html

Social Sciences

# *Do You Want to Know More?* Curiosity for Positive and Negative Life Stories of Ingroup and Outgroup Members

Afke van Egmond

*Supervisor*
Dr. Suzanne Oosterwijk (UvA)
*Reader*
Dr. Sennay Ghebreab (AUC)

**Abstract**

Current research on information seeking, an important aspect of curiosity, has shown that it can depend on context; for example, information seeking is different for positive as opposed to negative information, as well as about ingroup as opposed to outgroup members. The present study examines these relationships experimentally, using a paradigm developed for the purpose of this study. In this paradigm, participants were provided with pictures of ingroup and outgroup members, each with an emotional facial expression, along with a short description of this person's situation. These pictures and stories were either of positive or negative valence. Participants were constantly asked whether they wanted to know more about the person in the picture. The data was analysed using a generalised estimating equations (GEE) model. The results revealed that information seeking was higher for negative than for positive information, which can be explained by morbid curiosity. Overall, people did not seek information differently about ingroup than outgroup members. However, this changed when valence was taken into account: participants sought the least information about ingroup members in trials containing positive emotional information. This could be because this type of information is the least novel and uncertain. Future research should examine how the way people seek information influences what people learn about others.

Keywords and phrases: *information seeking, curiosity, outgroup bias, valence, empathy*

# I. Introduction

In today's society we are continuously confronted with the life stories of other people. We listen to the life events shared by family, friends, neighbours, and colleagues, whether they are part of the social we belong to, the ingroup, or the group we do not belong to, the outgroup We learn about other people's lives through news, books, movies, and shows. When we learn about life stories of others we are seeking information, which is an important behaviour resulting from a state of curiosity. Research indicates that this information seeking behaviour seems to differ depending on the social context and the valence of the information that is sought; for example, we might learn differently about people that are part of our own social group than people who are not. Since information seeking helps us keep track of our social environment, it is important that we learn about people in our own social group in the same way as people in other social groups, in order to keep a realistic view of the social environment. However, it seems that this does not always happen; outgroup bias, or generalised negative beliefs about outgroup members, exists in virtually all diverse, modern societies. Therefore, it is important to examine where this outgroup bias comes from, and whether it is related to information seeking and curiosity, because this knowledge can help in targeting problems such as the racist and discriminatory behaviour that may result from it. Information seeking is driven by intrinsic motivation; similarly, new research suggests that empathising with others is also driven by intrinsic motivation. We may empathise when we learn about others' life events, and it has been shown that empathy can help in reducing outgroup bias. The present study incorporates all these ideas and aims to answer the question: How is curiosity for the positive and negative life stories of other people influenced by social group status?

Defined as the drive to learn and explore, to seek new information, with the goal to expand knowledge, curiosity is what intrinsically motivates people to seek information about their (social) environment (Kashdan & Silvia, 2009). One of its components is social curiosity, which explains that people acquire information about other people in order to gain a better understanding of their social environment (Kashdan et al., 2018; Renner, 2006). Gottlieb, Oudeyer, Lopes and Baranes (2013) argue that since human social and cultural structures change so quickly, understanding this environment by seeking information in this context is evolutionarily adaptive. They argue that the motivation for information seeking is independent of a reward, and that learning alone is enough of a reinforcer to engage in the process of seeking information. Therefore, they conclude that information seeking is associated with a high degree of intrinsic motivation. Loewenstein (1994) explains this relationship between motivation and curiosity with his information gap hypothesis, which states that people are motivated to seek information when there is a discrepancy between what they know and what they want to know. They then seek out information in order to move from a state of uncertainty to one of certainty. People are intrinsically motivated to seek out information about others in order to learn more about their social environment and to resolve uncertainty.

When people are learning about their social environment, for example when they want to know more about the life stories of others, the way they seek information can differ depending on the context. For example, Baumeister, Bratslavsky, Finkenauer and Vohs (2001), as well as Taylor (1991), have shown that the processing of negative information is more ‘costly’ than positive information, meaning that it requires more cognitive resources. Based on this idea, it can be hypothesised that people are less likely to seek out negative information than positive information and would, therefore, prefer to learn more about positive life stories than negative ones. A contradicting idea comes from the concept of morbid curiosity, which describes a curiosity for intense, extremely negative information (Litman, 2005; Oosterwijk, 2017). Zuckerman (1990) argues that people seek out this type of negative information because they expect that it will induce a feeling of sensation. Baumeister et al. (2001) hypothesise that being sensitive to negative information is evolutionarily adaptive, which might also explain why people tend to seek out negative information (Oosterwijk, 2017). People respond to positive and negative information differently, and therefore information seeking in the context of learning about other’s positive and negative life stories may also differ, although the exact direction of this difference is not clear yet and should in the future be further investigated.

Information seeking may depend on the social context, in particular the interaction with different social groups. The social group that one belongs to is referred to as the ‘ingroup’, and the group that one does not belong to is called the ‘outgroup’ (Fiske, 1998; Tajfel, 1970). Tajfel and Turner (1979) argue that the sense of belonging to a group helps to build not only someone’s social identity, but also their self-image and self-esteem. For that reason, people tend to enhance the status of their ingroup in order to reinforce their self-image. Similarly, evolutionary psychologists speak of ingroup favouritism and outgroup hostility as adaptive features that promote survival (Masuda & Fu, 2015; Neuberg, Smith, & Asher, 2000; Van Vugt & Schaller, 2008). These ideas illustrate why people have different expectations of and assumptions about ingroup and outgroup members, which often persist because of confirmation bias, or the tendency to seek out information that confirms one’s own beliefs (Nelson, 2014). Aboud (1976) conducted an experiment to show that people indeed seek out different kinds of information about ingroup than outgroup members. She offers different explanations for why these discrepancies exist, the most significant one being that when people are communicating with members from other social groups, they seek information that affirms their expectations that these people are somehow different from themselves - effectively performing according to confirmation bias. Similarly, when people are exposed to information that does not agree with their beliefs, they disregard it. Uncertainty could be another reason why people might seek information differently for outgroup than ingroup members. As previously discussed, uncertainty drives curiosity (Gottlieb et al., 2013). It can be hypothesised that people are more uncertain about individuals from different social groups than their own, and therefore are more driven to learn about these people (Weary & Jacobson, 1997). In conclusion, seeking information about a person seems to depend on what social group that person belongs to in relation to one’s own.

Seeking out information about others, such as learning more about their life stories, might naturally activate empathy. Empathy is the cognitive process that helps people understand each other’s mental and emotional state, and in particular enables them to share this state (Avenanti, Bueti, Galati, & Aglioti, 2005; Stephan & Finlay, 1999). Until recently, the consensus was that empathising happened mostly automatically; evidence from multiple fields supported this claim. For instance, early research done by developmental psychologists Haviland and Lelwica (1987) showed that infants as young as ten weeks old mimicked their mothers’ emotions, and emotional mimicry is thought to be a predictor of empathy (Harrison, Morgan, & Critchley, 2010). Congruently, research from social psychology has robustly shown that emotional contagion, also an important building block of empathy, happens mostly automatically (Hatfield, Carpenter, & Rapson, 2014; Hatfield, Rapson, & Le, 2013). Neuroscientific research supports this idea on a neural level, by showing that performing an action is supported by similar neural systems as observing that same action (Gallese & Goldman, 2002). Zaki (2014) calls this parallel activation 'neural resonance', and he argues that this happens reflexively and even in the absence of attention.

Even though the previously discussed studies all seem to indicate that empathy is automatic, recent evidence has shown that empathy is also context-dependent. First of all, Decety, Yang and Cheng (2010) showed that in some cases, neural resonance is not completely automatic. They found that physicians, who are trained to 'turn off' empathy for their patients, had different early-latency event-related potentials than control participants when witnessing others being exposed to painful stimuli. This early latency indicates that the difference in the responses of physicians and control participants was rooted mostly in automaticity. In other words, physicians had less of an automatic empathy response; after years of down-regulating empathy for pain, this reaction simply did not unfold. Similarly, Sheng and Han (2012) measured participants' neural activity when observing ingroup and outgroup members experiencing pain, and found that this activity significantly differed in these two conditions. This effect again happened at such an early latency that it can be concluded that people empathise less automatically with outgroup members than ingroup members. Other studies using several different methods have supported the idea that people empathise less automatically with outgroup members than ingroup members (Cikara & Fiske, 2011; Cikara & Van Bavel, 2014; de Waal, 2008; Harris & Fiske, 2006). The contextual influence of empathy was also examined by Meffert, Gazzola, Den Boer, Bartels and Keysers (2013), who found that psychopaths have been shown to have abnormal neural activation in situations where empathy is expected to be spontaneously elicited, but also that this difference is reduced when they are explicitly asked to empathise.

Based on the evidence found by Meffert et al. (2013) regarding empathy in psychopaths, Keysers and Gazzola (2014) make a distinction between the capacity and the willingness to empathise. This suggests that empathy is not something that can only be activated automatically, but that willingness or motivation can play a role in it. Zaki (2014) also claims that since empathy is relatively resource-demanding, motivation is necessary to either approach or avoid empathising with others. Since processing negative emotions or information is more costly than positive information (Baumeister et al., 2001; Taylor, 1991), it is inherently implied that empathy can be more costly depending on the emotional context, and therefore might be avoided in some cases. Indeed, Cameron et al. (2016) as well as Zaki (2014) argue that empathy is an effortful process with a high cognitive cost. Based on this idea, Cameron, Cunningham, Saunders and Inzlicht (2018) state that empathising is a choice, based on the cost and benefits of its consequences. In short, we cannot take the automaticity of empathy for granted; the present study assumes that people do have some agency in empathising, in the form of motivation to engage in the resource-demanding process of empathy.

Based on these ideas, it seems that empathising with is similar to seeking information about someone in the sense that they both are driven by motivation, and can depend on context. Empathy is relevant for the process of learning about others' life stories; research on motivated empathy, as well as research on information seeking, indicate that learning about others' life stories might be different depending on the social group this other belongs to. Additionally, research on information seeking indicates that there might be a difference in positive and negative information seeking; this preference for either positive or negative information might even depend on the social group of the person someone is seeking information about. People might be more motivated to seek negative information about ingroup members than outgroup members, since the former are more relevant to them than the latter (Fiske, 1998; Tajfel, 1970), and processing negative information is more costly than positive information. Therefore, people might be less motivated to spend these extra resources necessary for negative information on the less relevant (the outgroup) than on what directly affects them (the ingroup). Conversely, based on the idea of confirmation bias, it is possible that people seek more positive information about ingroup members and more negative information about outgroup members, since this information agrees with people's expectations. The present study will examine these relationships and will aim to answer the question: How is curiosity for positive and negative life stories of other people influenced by social group status?

In order to measure how curiosity for life stories is influenced by the social context, the present study carried out an experiment, for the purpose of which a new paradigm was developed, which simulated real-life experiences of learning about others' life stories. In this paradigm, participants were shown images of ingroup members and outgroup members expressing positive (joy, surprise) and negative (sadness, fear) emotions, that were taken from the Amsterdam Dynamic Facial Expressions Set (ADFES; van der Schalk, Hawk, Fischer, & Doosje, 2011). This dataset contains pictures of North-European and Mediterranean models, which were validated by van der Schalk et al. (2011) to be the ingroup and outgroup, respectively, for Dutch participants. Along with these pictures, participants were given a one- sentence description of an emotional event, corresponding to the facial expression, that the person in the picture had supposedly experienced. Participants were then asked whether they wanted to know more about this person. Information seeking was quantified in this way, representing the number of times participants asked for more information in each trial.

The present study expected first of all that information seeking would be lower for negative than for positive information, based on the idea that processing negative information is more costly than processing positive information (Baumeister et al., 2001; Taylor, 1991). Additionally, it was expected that information seeking would be lower for outgroup members than for ingroup members, since outgroup members are less relevant, and studies have shown that people are have negative generalised assumptions about outgroup members (Fiske, 1998; Tajfel, 1970). Last of all, it was expected that the difference between information seeking for outgroup and ingroup members was bigger for negative than for positive information, because of the higher relevance of ingroup members as well as the cost of negative information processing.

The present study is relevant both in a societal as well as in an academic way. The societal relevance is evident: Prejudice exists as a result of (implicit) bias against outgroup members, and gaining a better understanding of where this bias comes from is key in trying to reduce its negative effects such as discriminatory and racist behaviours that manifest in diverse societies worldwide. Research has shown that empathy, or sharing others' experiences, can reduce racial bias (Drwecki, Moore, Ward, & Prkachin, 2011; Inzlicht, Gutsell, & Legault, 2012; Sheng & Han, 2012). If it is indeed true that people are less motivated to learn more about outgroup members than ingroup members, it can be concluded that motivation is what should be targeted in stimulating learning more about and thus sharing experiences of others, which can reduce outgroup bias. Furthermore, the academic relevance of the study is expressed in multiple ways. First of all, it is an exploratory study aiming to establish a relationship between concepts of curiosity, social identity, motivation, and empathy. In order to establish these links, validated questionnaires measuring concepts of curiosity (social as well as morbid) and empathy are included, so that correlations between these concepts and positive and negative information seeking can be examined. This is helpful in validating the present paradigm and will shed more light on how seeking information about other people's life stories is related to trait social curiosity and different aspects of trait empathy. Moreover, by putting these concepts in the same context for the first time in an innovative manner, the present study could make use of multiple insights from research about each of these concepts and link them to build different hypotheses. These hypotheses were tested using a paradigm that was newly developed for the purpose of the experiment. This paradigm mimics real-life situations, which increases the external validity of the present study; in addition, it can be used in future research, perhaps instead of paradigms that are currently used to measure concepts of information seeking or curiosity.

## II. Methodology

### Participants

Participants were recruited from the participant pool of the Psychology department at the Universiteit van Amsterdam (UvA; $N = 56$). The sample consisted of 12 males and 42 females; the average age of the sample was 19.3. Participants received credit for participating in the study. The Ethics Review Board of the Faculty of Social and Behavioural Sciences at the UvA reviewed the proposed study and gave permission to conduct the experiment.

## Design

The present study has a 2 (social group: $ingroup(0)/outgroup(1)$) by 2 (valence: $positive(0)/negative(1)$) design. Social group and valence are both within-subject variables. The dependent variable is the number of times participants asked for more information about the person in each picture. This variable is referred to as information seeking, and represents the number of times a subject asked for more information in one trial.

## Materials

The experiment was designed using the online survey tool Qualtrics. The pictures of ingroup and outgroup members displaying emotional facial expressions were selected from the validated ADFES (van der Schalk et al., 2011). This set contains stills from videos of both male and female models, showing one of nine emotions: joy, surprise, pride, anger, fear, sadness, disgust, contempt, and embarrassment. In a validation experiment, joy and surprise were rated as positive emotions, and fear and sadness were rated as the most negative (van der Schalk et al., 2011). Therefore, the present study chose to use these four emotions in order to avoid a confounding influence of valence. As previously mentioned, the models in the images of the ADFES are either North-European or Mediterranean; these groups are respectively defined as ingroup and outgroup for Dutch participants, as validated by van der Schalk et al. (2011).

The stories that were presented along with these pictures were situational and emotional (corresponding to the emotion on the picture), and were written for the purpose of this study. The topics of these stories were counterbalanced across ingroup and outgroup members of the same gender, to ensure that effects could not be driven by particular combinations between stories and models. Stereotype-evoking information, such as names or gendered pronouns, was avoided as much as possible during writing. The stories were split into nine parts that all carried equal amounts of emotional information. For example, the following two sentences, each carrying a substantial amount of emotional information, were subsequently used: “A few months ago, my sister suddenly passed away.”; “She was travelling, and she got into a big bus crash.”. For an example trial of the paradigm, see Appendix 1. All stories that were used (all in Dutch) are provided in Appendix 2.

## Procedure

Before the experiment started, participants read an information letter and gave informed consent. They were given a short instruction, in which they were told that the aim of the study was to examine how people responded to emotional information and pictures of certain emotional facial expressions. They received a short explanation of the paradigm and were told that after the paradigm they would have to fill in some questionnaires, following which the paradigm started. Participants were shown a picture of a person displaying a certain emotional facial expression, along with a short, one-sentence description of an emotionally evocative event that this person experienced. Some examples of events that were used are losing a family member, being promoted, getting into an accident, and having a child. The event was related to the emotion the model expressed in the picture. After viewing this picture and reading the description, participants were asked whether they wanted to know more about this person’s situation or not. If they did, they were given another sentence describing the person’s situation. Participants were given this choice eight times; after every new sentence they could choose to stop. When they stopped, or when they had been given eight new additions to the story, participants answered questions about the emotions the person in the picture was feeling. These questions consisted of a list of nine emotions, for each of which participants rated how strongly they felt that the model was experiencing that emotion, on a Likert scale from 1 (not at all) to 10 (very strong). This marked the end of one trial; the full paradigm consisted of 16 trials, showing a new picture with a new story each time.

Out of the 16 models, eight were ingroup (North-European) and eight were outgroup (Mediterranean). In each set of eight models, four were male and four were female. Because it is possible that ingroup/outgroup status interacted with the gender of the model, in terms of chosen information, gender was taken into account during the data analysis. Each of the four models in all gender groups displayed one of the four emotions that were chosen to be used; two with positive valence (joy, surprise), and two with a negative valence (fear, sadness). Subjects were randomly assigned to one of two counterbalance conditions so that the same stories were paired with pictures of ingroup members in one condition and outgroup members in the other, and vice versa. The order in which the sixteen trials were presented to participants was fully randomised.

After the main experiment was over, participants were asked to report some demographic information such as their age and gender. In addition, they were asked whether they identified as Dutch or non-Dutch, in order to define their ingroup and outgroup positioning. This was done because even though the survey was conducted in Dutch, it could not be assumed that all participants identified as being Dutch. Lastly, they were asked to fill in the Five-Dimensional Curiosity Scale (5-DC; Kashdan et al., 2018), the Interpersonal Reactivity Index (IRI; Davis, 1983), and the Morbid Curiosity in Daily Life scale (Oosterwijk, 2017). These questionnaires were included in order to obtain data on the effectiveness and validity of the paradigm itself, by finding correlations between each of the subscales of curiosity and empathy with information seeking, both positive and negative.

## Data Analysis

The data analysis of the experiment was done using IBM's SPSS statistical analysis programme. Because the dependent variable, information seeking, measures count data, normality cannot be assumed. The most appropriate analysis technique for count data in combination with a repeated measures design is a generalised estimating equations (GEE) model. GEE is an extension of generalized linear models (GLM) and is used to model the average response over a (sub)population or cluster, as opposed to subject-specific responses (Garson, 2013). It allows for non-normality of the dependent variable, as well as correlated data. In addition, GEE allows for more complex data analysis than GLM, such as repeated measures. In order to verify whether GEE was indeed the best model to use for this data, the normality of the dependent variable (information seeking) was tested. Measurements existed over the full range of information seeking, its minimum and maximum at the most extreme values of $0$ and $8$ respectively. Its overall mean was revealed to be $2.74$, which is relatively far from the middle value of $4$. This already suggested non-normality. Additionally, normality tests, normal Q-Q plots, as well as a histogram showing the frequency of information seeking, showed that the variable information seeking was not normally distributed. Interestingly, the histogram showed peaks at both extremes ($0$ and $8$). Therefore, it was concluded that GEE was indeed a good fit for this data.

In GEE, there are different types of models that can be used; the present study used a repeated measures Poisson log-linear model. This model was chosen since a Poisson distribution with a log-linear link function is known to best approach the distribution of count data (Garson, 2013). The type of correlation matrix that was assumed was chosen by trying out different types of correlation matrices that SPSS offers in GEE and checking the goodness-of-fit of each model. In GEE, the goodness-of-fit is approached by the quasi-likelihood under independence criterion (QIC). The correlation matrix assumed in the model with the best fit, and thus lowest QIC, was used in the final model. When testing these correlation matrices, a robust estimator was used, since this is robust against choosing the wrong correlation matrix and is generally used when the correct correlation matrix has not yet been selected. However, after the correct correlation matrix had been identified, a model-based estimator was used. This estimator does assume a correct selection of the correlation matrix, and works better for smaller sample sizes.

Two models were tested regarding social group: first of all, all participants were included ($N = 56$), and secondly, the participants who indicated that they identified as non-Dutch were filtered out so that $N = 46$, in order to examine whether these participants being included in the first model influenced the results. In the second model, the significance of all three effects that were tested for did not change; excluding participants who identified as non-Dutch did not change the results significantly. Since van der Schalk et al. (2011) used only participants that identified as Dutch to validate the difference between ingroup and outgroup members in the pictures in the ADFES, the present study decided to use this same assumption and exclude all participants that identified as non-Dutch.

This final model was used to test for the main effects of both social group and valence, and an interaction effect of social group and valence on information seeking. In GEE, Wald chi-square ($W^2$) is used as a test statistic; the bigger Wald chi-square, the stronger the effect. For significance, an alpha of $0.05$ was used.

Besides analysing data obtained by the experiment, the questionnaires that were included were used to examine correlations between different personality traits and positive and negative information seeking. The 5-DC calculates the five dimensions of curiosity: joyous exploration, deprivation sensitivity, stress tolerance, social curiosity, and thrill seeking. In the present paper, only correlations with social curiosity are reported; Cronbach's alpha for the social curiosity scale was $\alpha = 0.787$. Correlations with the IRI are reported for all its four subscales of empathy: fantasy ($\alpha = 0.789$), empathic concern ($\alpha = 0.702$), perspective taking ($\alpha = 0.556$), and personal distress ($\alpha = 0.777$). The Morbid Curiosity in Daily Life is a single scale, which had a Cronbach's alpha of $\alpha = 0.601$; this scale was also correlated with positive and negative information seeking. Lastly, in order to establish the effectiveness of the paradigm, some statistics about the way participants responded across trials were retrieved.

### Hypotheses

The present study expected that the main effect of social group on information seeking would be significant and negative, meaning that participants would want to know more about ingroup members than outgroup members (H1). The main effect of valence on information seeking was expected to be significant and positive. In other words, the present study expected that participants would want more information about positive than negative stories (H2). Additionally, a significant interaction between social group and valence on information seeking was expected; the present study hypothesised that the difference between information seeking for ingroup and outgroup members was bigger for negative than for positive information (H3). Lastly, the present study expects correlations (positive, negative, or both) between information seeking, social curiosity, and morbid curiosity. Additionally, a correlation (again positive, negative, or both) of subscales of empathy with information seeking, is expected.

## III. Results

### Social Group

Contrary to H1, the analysis showed that social group did not have a significant effect on information seeking ($W^2 = 1.837$; $p = 0.175$).

### Valence

As predicted, the analysis demonstrated a significant effect of valence ($W^2 = 13.107$; $p = 0.000$). However, in contrast to H2, participants wanted to know more about negative ($M = 2.99$; $SE = 1.181$) than positive ($M = 2.64$; $SE = 0.170$) life stories.

### Social Group*Valence

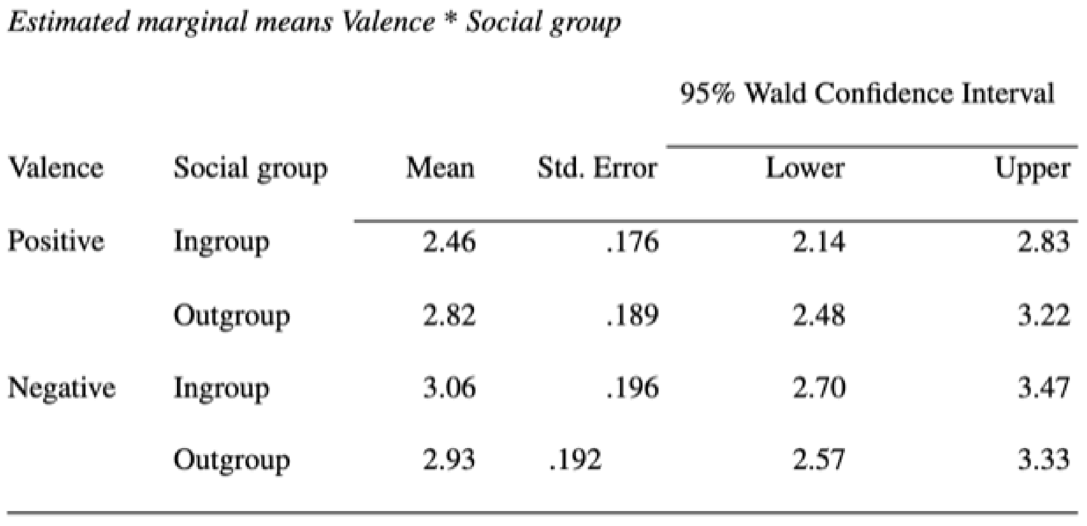
Congruent with H3, an interaction between valence and social group on information seeking was found ($W^2 = 6.800$; $p = 0.009$). Pairwise comparisons, shown in Table 1, revealed that for positive information about ingroup members, information seeking was significantly lower than all other combinations of social group and valence. The other groups did not significantly differ among each other. These significant relationships are presented in Figure 1 by use of asterisks. In conclusion, like Hypothesis 3 predicted, social group and valence do interact, but not in the way that was hypothesised.

### Correlations

In order to find out whether the paradigm is a good measure for social curiosity, participants' scores on the 5-DC social curiosity subscale (Kashdan et al., 2018) were correlated with the positive and negative information seeking score. Positive information seeking was not significantly correlated with social curiosity ($r = 0.161$; $p = 0.240$), but there was a significant positive correlation between negative information seeking and social curiosity ($r = 0.278$; $p = 0.040$). In other words, the higher the participants scored on social curiosity, the higher their scores for negative information seeking were.

In exploratory fashion, the four subscales of the IRI were correlated with positive and negative information seeking. The IRI's subscale of perspective

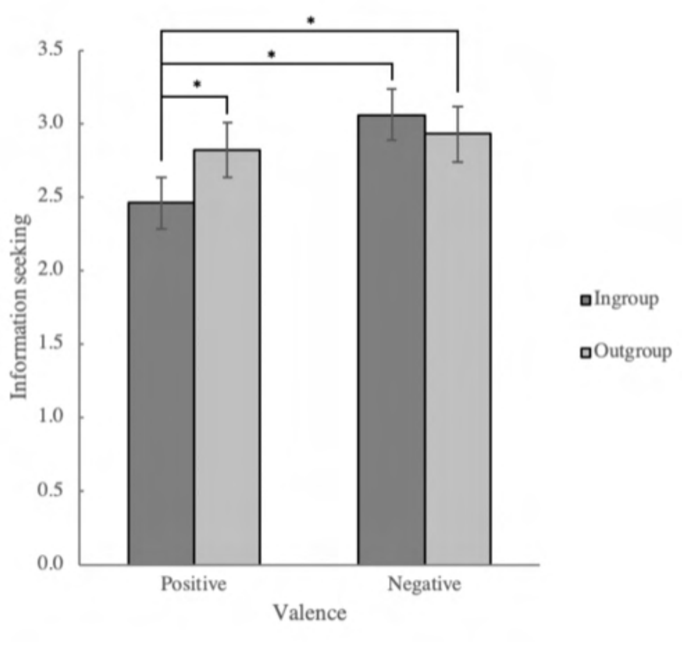
**Table 1.**



*Estimated marginal means Valence * Social group*

| Valence | Social group | Mean | Std. Error | 95% Wald Confidence Interval Lower | Upper |
|---|---|---|---|---|---|
| Positive | Ingroup | 2.46 | .176 | 2.14 | 2.83 |
| | Outgroup | 2.82 | .189 | 2.48 | 3.22 |
| Negative | Ingroup | 3.06 | .196 | 2.70 | 3.47 |
| | Outgroup | 2.93 | .192 | 2.57 | 3.33 |

**Figure 1**. Information seeking for Valence*Social group.



taking was positively correlated with positive information seeking ($r = 0.362$, $p = 0.035$), but not with negative information seeking ($r = 0.199$, $p = 0.154$). The three other subscales did not significantly correlate with either positive or negative information seeking (all $p > 0.05$).

Lastly, the Morbid Curiosity in Daily Life scale was correlated with positive and negative information seeking, but no significant correlations were found ($r = 0.064, p = 0.642; r = 0.055, p = 0.696$, respectively).

## Paradigm

To establish the effectiveness of the paradigm that the present study used, the way participants responded across trials was examined. All participants ($N = 56$) were taken into account during this analysis. There were three participants who chose the maximum available information for all sixteen trials; four participants never chose to get more information in any of the sixteen trials. The rest of the participants' ($N = 49$) scores for information seeking varied across trials. From this it seems that the majority of the participants reacted differently to different pictures and stories, which is congruent with the intention of the paradigm.

## IV. Discussion and Conclusion

The results of the statistical analyses were compared to the hypotheses. First of all, H1 stated that a negative effect of social group on information seeking was expected. This effect was not found, meaning that there was no significant difference in the number of times participants asked for more information about ingroup members than outgroup members. This might be explained by demand characteristics playing a role in the current experiment; at least some participants might have guessed the true purpose of the study, which might have affected their responses. Additionally, gender might have been a confounding variable in this effect. The present study studied ingroup and outgroup in terms of (appeared) ethnicity; however, one's ingroup and outgroup may also be defined by gender. Since the present sample was $78\%$ female, but the trials were $50\%$ female, the two-dimensionality of social group and the fact that the sample was predominantly female together may have confounded the results and lead to a diminished effect of social group on information seeking. Lastly, since the present sample consisted of psychology students, and thus they had all studied psychological concepts such as stereotypes and outgroup bias, they might have been more aware of their bias. Awareness of outgroup bias has been shown to reduce the bias itself (Kawakami, Dion, & Dovidio, 1998; Pope, Price, & Wolfers, 2018); therefore, the results of the present study might be biased by these characteristics of the sample.

A significant effect of valence on social group was found, however, not in the form that H2 predicted. Even though processing negative information is more costly, participants in the present study chose to get more information in trials containing negative emotional pictures and stories than in positive trials. This effect may be explained by morbid curiosity - the seeking of negative information because of its associated feeling of sensation (Zuckerman, 1990) or higher informational value (Baumeister et al., 2001; Oosterwijk, 2017). This curiosity seems to be stronger than the tendency to avoid processing negative information because of its higher cost. In fact, the paradigm created for the present study may be better at measuring morbid curiosity than paradigm currently used by Oosterwijk (2017). In that paradigm, participants are provided with a description of a situation and are asked whether they want to see the picture that belongs to the description. In this case, they have to choose between looking at a picture or a white screen, both of which take the same amount of time. In the case of negative emotional pictures, it makes sense that participants sometimes choose not to look at this picture, and instead look at a white screen. However, in the case of positive emotional pictures, participants may prefer looking at a picture rather than at a white screen for several seconds, and thus their default response may be to say yes. In other words, when participants have to invest the same amount of time into either choice, even when they are not particularly interested in the picture, it may still be preferable to see the picture over an empty screen, but only when the information is positive. As a result, research with this paradigm has found that people choose positive information more often than negative information (Oosterwijk et al., in prep.). In the current paradigm, however, choosing not to get more information saves participants time. Therefore, the default choice may be not to get additional information. Increased information seeking for negative information as opposed to positive information in the present study happened against the cost of being emotionally disturbed by the negative information, and against the cost of time investment necessary if participants wanted to know more. In this sense, the presently developed paradigm may be a better measure of morbid curiosity than the currently used paradigm.

However, it should be noted that no correlations were found between information seeking, positive or negative, and the Morbid Curiosity in Daily Life scale. This might be related to the scale's relatively low Cronbach's alpha ($\alpha = 0.601$); the scale may not have properly measured the concept of morbid curiosity in the present study. It is also good to note that the stimuli used in both the present paradigm and the paradigm for morbid curiosity used by Oosterwijk (2017) are very different. The present study used negative life stories about people, which contained different information than the pictures shown in the paradigm used by Oosterwijk (2017), which portray scenes involving violence and death. Therefore, the tendency to seek negative information found in the present study may not reflect the same concept as measured in the morbid curiosity paradigm. Although the construct that the present study measured (which we may call negative social curiosity) is likely to be related to morbid curiosity, it is also different, since the negativity of visual scenes involving violence and death cannot be compared to the negative valence of the life stories of the present study. Therefore it is concluded that negative social curiosity should not be measured using a morbid curiosity scale. It is good to take this distinction into account in case the present paradigm is used in future research.

Even though no main effect of social group in information seeking was found, there was an interaction effect between social group and valence on information seeking, which agrees with H3. However, the specific interaction that was found was different than the one H3 had predicted; participants wanted to know the least in trials that contained positive information about ingroup members, than in all other trials with other combinations of social group and valence. This might be explained by the fact that curiosity is often elicited by novelty and uncertainty (Kashdan & Silvia, 2009; Kashdan et al., 2018; Litman, 2005). Ingroup members are who we know most about and thus are least uncertain about, since they are part of our own social group (Zebrowitz, Bronstad, & Lee, 2007). Conversely, we are more uncertain about outgroup members, and specifically positive information about them may be the most novel, since we are often biased to think negatively about outgroup members (Masuda & Fu, 2015; Neuberg et al., 2000; Tajfel & Turner, 1979; Van Vugt & Schaller, 2008). This uncertainty and novelty regarding outgroup members and especially positive information about them may explain the present findings.

The lack of a main effect of social group on information seeking does not explain where outgroup bias comes from; the expected finding that participants wanted to know less about outgroup members than ingroup members could have suggested that people are biased against learning new information about people in different social groups than our own. On the contrary, the present study found no difference in information seeking about ingroup and outgroup members. This finding poses a positive outlook on problems of outgroup bias: The participants of the present study did want to learn about outgroup members, and they wanted to know more about them regardless of the valence of the information they received. This might mean that they are not biased against outgroup members. However, this conclusion cannot be drawn with certainty, since the present study did not measure participants' bias.

Future research should include a measure of outgroup bias in order to establish whether learning about outgroup members as much as ingroup members is correlated with a reduced outgroup bias. In addition, studies have shown that empathy can reduce racial bias (Cameron et al., 2016; Drwecki et al., 2011; Sheng & Han, 2012); future research should try to determine whether learning more about an outgroup member also changes one's view of that person and possibly reduces bias. For example, the paradigm could be adapted to include personal judgement scales of the person whose picture was seen. Then, data could be analysed to find out whether learning more about someone changes one's judgement of that person. Another interesting way to look at the effect of learning more about ingroup and outgroup members is to utilize an implicit bias task about people that participants did and did not learn about. Alternatively, a memory task could be used to find out how information learned about ingroup and outgroup members is remembered. If participants remember information about outgroup members better, this may indicate that they do process this information differently, which could give an insight into the underlying mechanism of outgroup bias. Adding to knowledge about what causes and what reduces outgroup bias is of tremendous importance in attempting to change (often implicit) racist perspectives adopted by people in many diverse societies of today.

Besides examining the effects of social group and valence on information seeking, several scales were included in order to examine whether the concepts that these scales measured were correlated with information seeking in this paradigm. The social curiosity subscale of the 5-DC measures the tendency to acquire information about the social environment; people who score high on this scale have a strong interest in other people's thoughts and actions. It was expected that social curiosity was correlated with both positive and negative information seeking; people who score high on the personality trait social curiosity, should choose to know more about other people's life stories. However, there was only a significant correlation between the social curiosity subscale of the 5-DC and negative information seeking; the higher participants scored on social curiosity, the more often they wanted to know more about other people's negative life events, but this relationship was not found for positive information. It is worth mentioning that Kashdan et al., (2018), in their study validating the 5-DC, found that their subscale of social curiosity correlated strongly with a tendency to gossip. Hartung and Renner (2013) state that social curiosity and gossip are related concepts, but that social curiosity serves a function of gathering information about the social environment, whereas gossiping is more sensational, and done for entertainment purposes. Since Kashdan et al. (2018) found a correlation between their social curiosity subscale and a tendency to gossip, it might be that this subscale is in fact a better measure of the construct of gossip than of social curiosity. The fact that gossip often involves discussing negative information about others (Feinberg, Willer, Stellar, & Keltner, 2012) could then explain why the present study found a correlation between the social curiosity subscale of the 5-DC and negative information seeking.

Alongside the 5-DC, the IRI was included in the present experiment. The IRI consists of items measuring four distinct aspects of empathy: fantasy, perspective taking, empathic concern, and personal distress. A correlation was found with the subscale of perspective taking and positive information seeking, but not with negative information seeking. This means that people who score high on perspective taking, or in other words often spontaneously adopt other people's point of view (Davis, 1983), had a tendency to ask for more positive information, but not necessarily negative information. It can be said that because these people spontaneously experience someone else's perspective, they enjoy being provided with positive information, since they then empathise with and therefore experience the positive emotions associated with the information (Batson, 2013). However, one has to be careful with drawing this conclusion, since Cronbach's alpha of the perspective taking subscale was low ($\alpha = 0.556$), and therefore this subscale may not have accurately measured this construct in the present study. The other three subscales of empathy (fantasy, emotional concern and personal distress) did not significantly correlate with information seeking, positive or negative, which suggests that participants' behaviour in the present paradigm does not reflect these aspects of empathy.

In conclusion, the present study is an exploratory one, and developed an innovative paradigm in order to establish a relationship between information seeking, social group, valence and empathy. It was found that participants preferred negative emotional information over positive, and that they were least curious of positive information about ingroup members. Because of the exploratory nature of the present study, it is important that it is replicated in order to verify that similar results are found across studies. If they are, these results greatly add to current knowledge in the field of each of these concepts and open up many new directions for future research. Future studies can elaborate on the results found in the present study, in order to more strongly establish how people seek information and how their way of seeking information changes their knowledge about the world and opinions about others. Relating this knowledge to outgroup bias will be incredibly important in reducing this bias and thus reduce the racism and discrimination that come as its result.

## Works Cited

Aboud, F. E. (1976). Self-evaluation: Information seeking strategies for interethnic social comparisons. *Journal of Cross-Cultural Psychology*, 7(3), 289–300.

Avenanti, A., Bueti, D., Galati, G., & Aglioti, S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. *Nature Neuroscience*, 8(7), 955– 960. https://doi.org/10.1038/nn1481

Batson, C. D. (2013). These things called empathy: Eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The Social Neuroscience of Empathy*. Cambridge, MA, US: MIT Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Cameron, C. D., Cunningham, W., Saunders, B., & Inzlicht, M. (2018). The ends of empathy:

Constructing empathy from value-based choice. *Psyarxiv.Com.* Retrieved from https://psyarxiv.com/d99bp/download/?format=pdf

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Hadjiandreou, E., & Inzlicht, M. (2016). Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Psyarxiv.Com.* https://doi.org/10.31234/osf.io/jkc4n

Cikara, M., & Fiske, S. T. (2011). Bounded empathy: Neural responses to outgroup targets' (mis)fortunes. *Journal of Cognitive Neuroscience*, 23(12), 3791–3803. https://doi.org/10.1162/jocn_a_00069

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science*, 9(3), 245–274. https://doi.org/10.1177/1745691614527464

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

de Waal, F. B. M. (2008). Putting the altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, 59, 279–300. Retrieved from https://www.annualreviews.org/doi/abs/10.1146/annurev.psych.59.103006.093625

Decety, J., Yang, C. Y., & Cheng, Y. (2010). Physicians down-regulate their pain empathy response: An event-related brain potential study. *NeuroImage*, 50(4), 1676–1682. https://doi.org/10.1016/j.neuroimage.2010.01.025

Drwecki, B. B., Moore, C. F., Ward, S. E., & Prkachin, K. M. (2011). Reducing racial disparities in pain treatment: The role of empathy and perspective-taking. *Pain*, 152(5), 1001–1006. https://doi.org/10.1016/j.pain.2010.12.005

Feinberg, M., Willer, R., Stellar, J., & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*, 102(5), 1015–1030. https://doi.org/10.1037/a0026650

Fiske, S. T. (1998). Stereotyping, prejudice and discrimination. In *The handbook of social psychology* (pp. 357–411).

Gallese, V., & Goldman, A. (2002). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12), 493–501. https://doi.org/10.1016/s1364-6613(98)01262-5

Garson, G. D. (2013). *Generalized linear models/generalized estimating equations*. Statistical Associates Publishing.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11), 585–593. https://doi.org/10.1016/j.tics.2013.09.001

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science*, 17(10), 847–853. Retrieved from https://journals.sagepub.com/doi/abs/10.1111/j.1467-9280.2006.01793.x?casa_token=qnpQXh6am_0AAAAA:8zXXQQzT9r0A81-fin9BoocTdOMu-LBTP6j c7dziyXZ33Bajb5sUEOyNxH59OF0CCc5C8rG
bw

Harrison, N. A., Morgan, R., & Critchley, H. D. (2010). From facial mimicry to emotional empathy: A role for norepinephrine? *Social Neuroscience*, 5(4), 393–400. https://doi.org/10.1080/17470911003656330

Hartung, F. M., & Renner, B. (2013). Social curiosity and gossip: Related but different drives of social functioning. *PLoS ONE*, 8(7). https://doi.org/10.1371/journal.pone.0069996

Hatfield, E., Carpenter, M., & Rapson, R. L. (2014). Emotional contagion as a precursor to collective emotions. In *Collective Emotions* (pp. 108–122). https://doi.org/10.1093/acprof:oso/9780199659180.003.0008

Hatfield, E., Rapson, R. L., & Le, Y.-C. L. (2013). Emotional contagion and empathy. In J. Decety & W. Ickes (Eds.), *The Social Neuroscience of Empathy*. Cambridge, MA, US: MIT Press. https://doi.org/10.7551/mitpress/9780262012973.003.0003

Haviland, J. M., & Lelwica, M. (1987). The induced affect response: 10-week-old infants' responses to three emotion expressions. *Developmental Psychology*, 23(1), 97–104. https://doi.org/10.1037/0012-

1649.23.1.97

Inzlicht, M., Gutsell, J. N., & Legault, L. (2012). Mimicry reduces racial prejudice. *Journal of Experimental Social Psychology*, 48(1), 361–365. https://doi.org/10.1016/j.jesp.2011.06.007

Kashdan, T. B., & Silvia, P. J. (2009). Curiosity and interest: The benefits of thriving on novelty and challenge. In S. J. Lopez & C. R. Snyder (Eds.), *The Oxford Handbook of Positive Psychology* (pp. 367–374). New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195187243.013.0034

Kashdan, T. B., Stiksma, M. C., Disabato, D. J., McKnight, P. E., Bekier, J., Kaji, J., & Lazarus, R. (2018). The five-dimensional curiosity scale: Capturing the bandwidth of curiosity and identifying four unique subgroups of curious people. *Journal of Research in Personality*, 73, 130–149. https://doi.org/10.1016/j.jrp.2017.11.011

Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. *Personality and Social Psychology Bulletin*, 24(4), 407–416. https://doi.org/10.1177/0146167298244007

Keysers, C., & Gazzola, V. (2014). Dissociating the ability and propensity for empathy. *Trends in Cognitive Sciences*, 18(4), 163–166. https://doi.org/10.1016/j.tics.2013.12.011

Litman, J. A. (2005). Curiosity and the pleasures of learning: Wanting and liking new information. *Cognition and Emotion* , 19(6), 793–814. https://doi.org/10.1080/02699930541000101

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*. https://doi.org/10.1037//0033-2909.116.1.75

Masuda, N., & Fu, F. (2015). Evolutionary models of in-group favoritism. *F1000Prime Reports*, 7(March), 1–12. https://doi.org/10.12703/P7-27

Meffert, H., Gazzola, V., Den Boer, J. A., Bartels, A. A. J., & Keysers, C. (2013). Reduced spontaneous but relatively normal deliberate vicarious representations in psychopathy. *Brain*, 136(8), 2550–2562. https://doi.org/10.1093/brain/awt190

Nelson, J. A. (2014). The power of stereotyping and confirmation bias to overwhelm accurate assessment: the case of economics, gender, and risk aversion. *Journal of Economic Methodology*, 21(3), 211–231. https://doi.org/10.1080/1350178X.2014.939691

Neuberg, S. L., Smith, D. M., & Asher, T. (2000). Why people stigmatize: Toward a biocultural framework. In *The social psychology of stigma.*

Oosterwijk, S. (2017). Choosing the negative: A behavioral demonstration of morbid curiosity. *PLoS ONE*, 12(7), 1–20. https://doi.org/10.1371/journal.pone.0178399

Oosterwijk, S., Snoek, L., Te Koppele, J., Engelbert, L., & Scholte, H. S. (in prep.). Choosing to view morbid stimuli involves reward circuitry.

Pope, D. G., Price, J., & Wolfers, J. (2018). Awareness reduces racial bias. *Management Science*, 64(11), 4988–4995.

Renner, B. (2006). Curiosity about people: The development of a social curiosity measure in adults. *Journal of Personality Assessment*, 87(3), 305–316. https://doi.org/10.1207/s15327752jpa8703

Sheng, F., & Han, S. (2012). Manipulations of cognitive strategies and intergroup relationships reduce the racial bias in empathic neural responses. *NeuroImage*, 61, 786–797. Retrieved from https://www.sciencedirect.com/science/article/pii/S1053811912004247

Stephan, W. G., & Finlay, K. (1999). The role of empathy in improving intergroup relations. *Journal of Social Issues*, 55(4), 729–743. https://doi.org/10.1111/0022-4537.00144

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 223(5), 96–102. https://doi.org/10.1038/scientificamerican1170-96

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In *The Social Psychology of Intergroup Relations*. https://doi.org/10.1016/S0065-2601(05)37005-5

Taylor, S. (1991). Asymmetrical effects of positive and negative events. *Psychological Bulletin*, 110(l), 67–85. Retrieved from https://taylorlab.psych.ucla.edu/wpcontent/uploads/sites/5/2014/10/1991_Asymmetrical-Effects_Positive«_Negative-Events_Mobilization-Minimization-Hypothesis.pdf

van der Schalk, J., Hawk, S. T., Fischer, A. H., & Doosje, B. (2011). Moving faces, looking places: Validation of the Amsterdam Dynamic Facial Expression Set (ADFES). *Emotion*, 11(4), 907–920. https://doi.org/10.1037/a0023853

Van Vugt, M., & Schaller, M. (2008). Evolutionary approaches to group dynamics: An introduction. *Group Dynamics*, 12(1), 1–6. https://doi.org/10.1037/1089-2699.12.1.1

Weary, G., & Jacobson, J. A. (1997). Causal uncertainty beliefs and diagnostic information seeking. *Journal of Personality and Social Psychology*, 73(4), 839–848. https://doi.org/10.1037/0022-3514.73.4.839

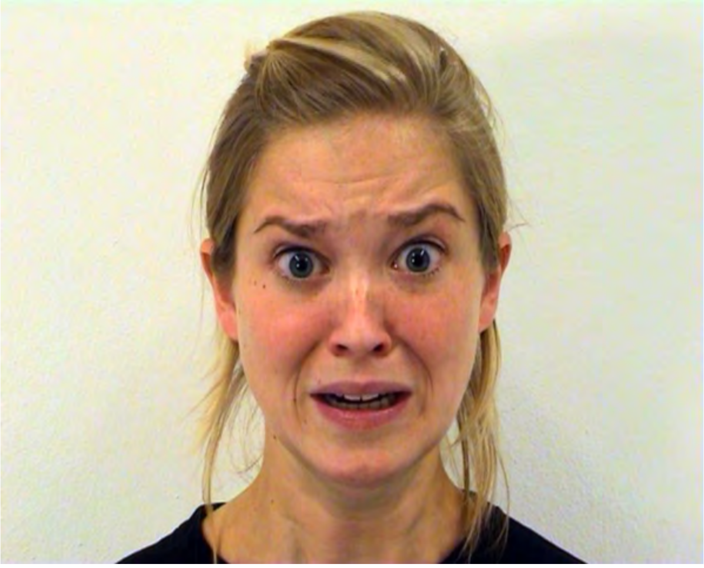
Zaki, J. (2014). Empathy: A motivated account. *Psychological Bulletin*, 140(6), 1608–1647. https://doi.org/10.1037/a0037679

Zebrowitz, L. A., Bronstad, P. M., & Lee, H. K. (2007). The contribution of face familiarity to ingroup favoritism and stereotyping. *Social Cognition*, 25(2), 306–338. https://doi.org/10.1521/soco.2007.25.2.306

Zuckerman, M. (1990). The psychophysiology of sensation seeking. *Journal of Personality*. https://doi.org/10.1111/j.1467-6494.1990.tb00918.x

## Appendix A

**Example of the stimuli material used.**



Een half jaar geleden ben ik ernstig gewond geraakt in een verkeersongeluk.
Wil je meer weten?

- Ja
- Nee

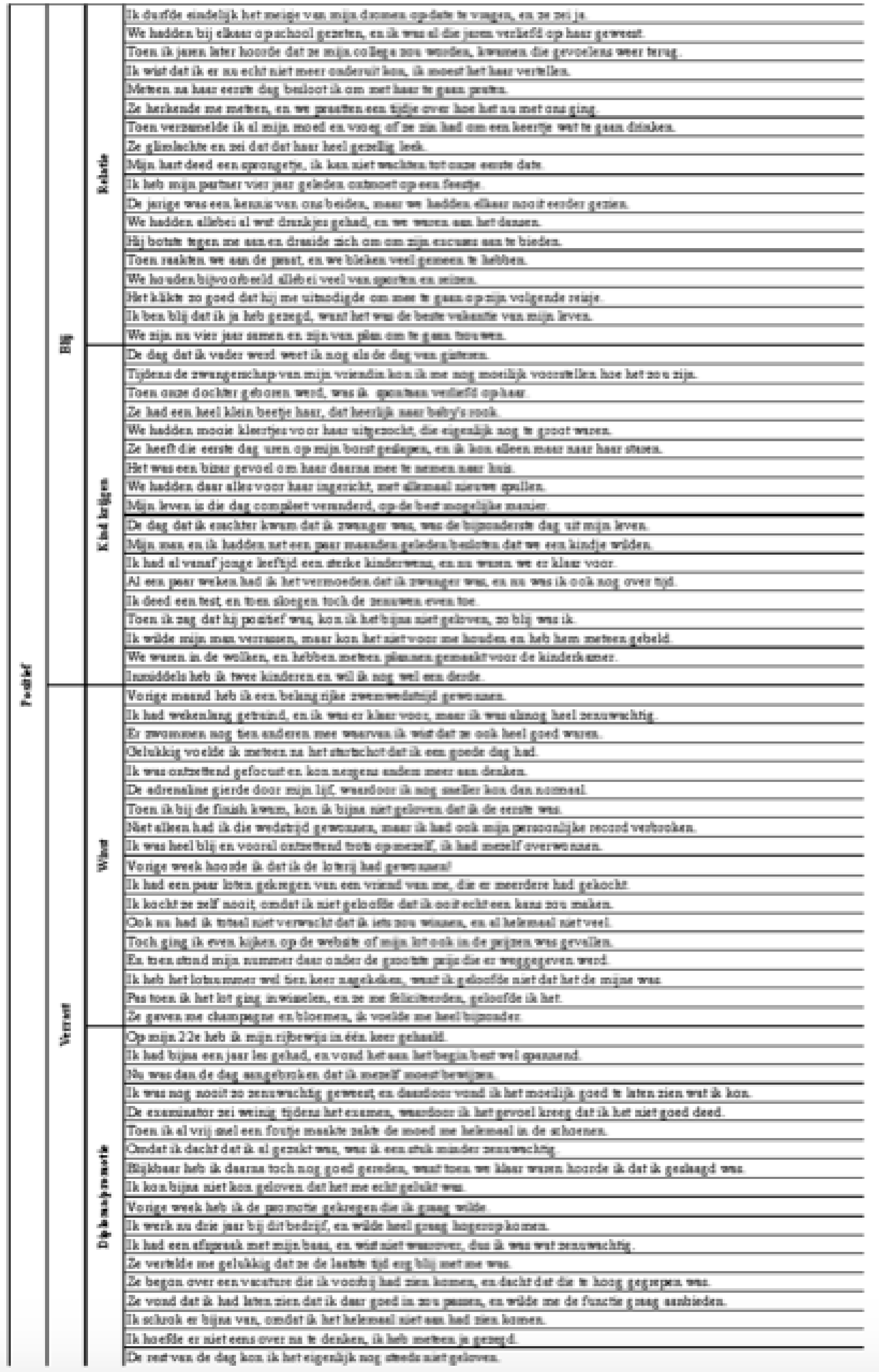
*Additional lines:*

1. Het was ochtend, het regende, en ik zat op de fiets onderweg naar mijn werk.
2. Op een kruising heeft een auto die door rood reed me geschept.
3. Ik kan me nog goed herinneren hoe de auto eruit zag: het was een blauwe Opel.
4. Het ging heel erg snel, ik realiseerde het me pas toen ik op de grond lag.

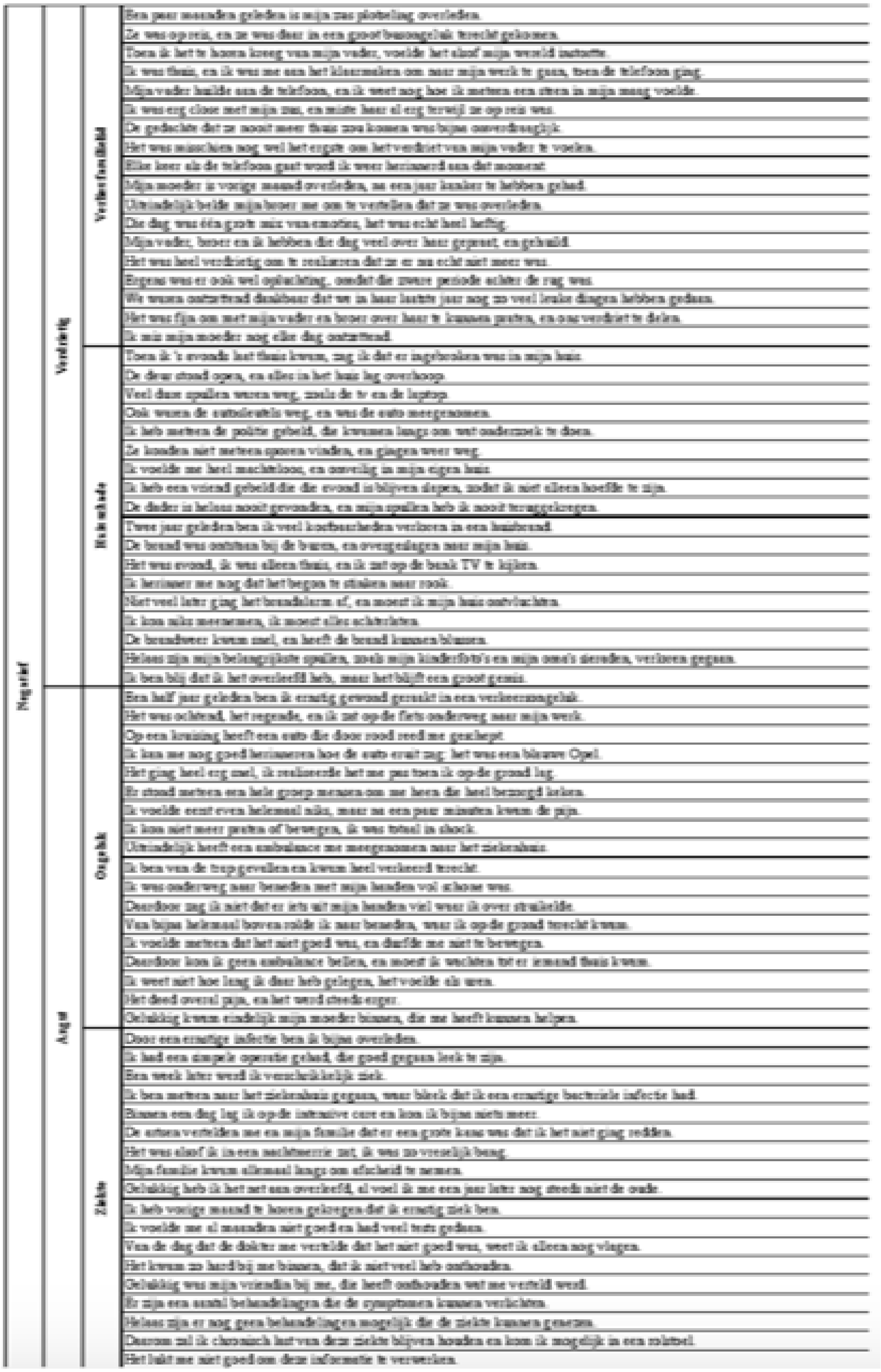
5. Er stond meteen een hele groep mensen om me heen die heel bezorgd keken.
6. Ik voelde eerst even helemaal niks, maar na een paar minuten kwam de pijn.
7. Ik kon niet meer praten of bewegen, ik was totaal in shock.
8. Uiteindelijk heeft een ambulance me meegenomen naar het ziekenhuis.

# Appendix B

**Emotional stories written for and used in the present study (in Dutch only).**



| | | | |
|---|---|---|---|
| Positief | Blij | Relatie | Ik durfde eindelijk het meisje van mijn dromen op date te vragen, en ze zei ja. |
| | | | We hadden bij elkaar op school gezeten, en ik was al die jaren verliefd op haar geweest. |
| | | | Toen ik jaren later hoorde dat ze mijn collega zou worden, kwamen die gevoelens weer terug. |
| | | | Ik wist dat ik er nu echt niet meer onderuit kon, ik moest het haar vertellen. |
| | | | Meteen na haar eerste dag besloot ik om met haar te gaan praten. |
| | | | Ze herkende me meteen, en we praatten een tijdje over hoe het nu met ons ging. |
| | | | Toen verzamelde ik al mijn moed en vroeg of ze zin had om een keertje wat te gaan drinken. |
| | | | Ze glimlachte en zei dat dat haar heel gezellig leek. |
| | | | Mijn hart deed een sprongetje, ik kan niet wachten tot onze eerste date. |
| | | | Ik heb mijn partner vier jaar geleden ontmoet op een feestje. |
| | | | De jarige was een kennis van ons beiden, maar we hadden elkaar nooit eerder gezien. |
| | | | We hadden allebei al wat drankjes gehad, en we waren aan het dansen. |
| | | | Hij botste tegen me aan en draaide zich om om zijn excuses aan te bieden. |
| | | | Toen raakten we aan de praat, en we bleken veel gemeen te hebben. |
| | | | We houden bijvoorbeeld allebei veel van sporten en reizen. |
| | | | Het klikte zo goed dat hij me uitnodigde om mee te gaan op zijn volgende reisje. |
| | | | Ik ben blij dat ik ja heb gezegd, want het was de beste vakantie van mijn leven. |
| | | | We zijn nu vier jaar samen en zijn van plan om te gaan trouwen. |
| | | Kind krijgen | De dag dat ik vader werd weet ik nog als de dag van gisteren. |
| | | | Tijdens de zwangerschap van mijn vriendin kon ik me nog moeilijk voorstellen hoe het zou zijn. |
| | | | Toen onze dochter geboren werd, was ik spontaan verliefd op haar. |
| | | | Ze had een heel klein beetje haar, dat heerlijk naar baby's rook. |
| | | | We hadden mooie kleertjes voor haar uitgezocht, die eigenlijk nog te groot waren. |
| | | | Ze heeft die eerste dag uren op mijn borst geslapen, en ik kon alleen maar naar haar staren. |
| | | | Het was een bizar gevoel om haar daarna mee te nemen naar huis. |
| | | | We hadden daar alles voor haar ingericht, met allemaal nieuwe spullen. |
| | | | Mijn leven is die dag compleet veranderd, op de best mogelijke manier. |
| | | | De dag dat ik erachter kwam dat ik zwanger was, was de bijzonderste dag uit mijn leven. |
| | | | Mijn man en ik hadden net een paar maanden geleden besloten dat we een kindje wilden. |
| | | | Ik had al vanaf jonge leeftijd een sterke kinderwens, en nu waren we er klaar voor. |
| | | | Al een paar weken had ik het vermoeden dat ik zwanger was, en nu was ik ook nog over tijd. |
| | | | Ik deed een test, en toen sloegen toch de zenuwen even toe. |
| | | | Toen ik zag dat hij positief was, kon ik het bijna niet geloven, zo blij was ik. |
| | | | Ik wilde mijn man verrassen, maar kon het niet voor me houden en heb hem meteen gebeld. |
| | | | We waren in de wolken, en hebben meteen plannen gemaakt voor de kinderkamer. |
| | | | Inmiddels heb ik twee kinderen en wil ik nog wel een derde. |
| | Verrast | Winst | Vorige maand heb ik een belangrijke zwemwedstrijd gewonnen. |
| | | | Ik had wekenlang getraind, en ik was er klaar voor, maar ik was alsnog heel zenuwachtig. |
| | | | Er zwommen nog tien anderen mee waarvan ik wist dat ze ook heel goed waren. |
| | | | Gelukkig voelde ik meteen na het startschot dat ik een goede dag had. |
| | | | Ik was ontzettend gefocust en kon nergens anders meer aan denken. |
| | | | De adrenaline gierde door mijn lijf, waardoor ik nog sneller kon dan normaal. |
| | | | Toen ik bij de finish kwam, kon ik bijna niet geloven dat ik de eerste was. |
| | | | Niet alleen had ik die wedstrijd gewonnen, maar ik had ook mijn persoonlijke record verbroken. |
| | | | Ik was heel blij en vooral ontzettend trots op mezelf, ik had mezelf overwonnen. |
| | | | Vorige week hoorde ik dat ik de loterij had gewonnen! |
| | | | Ik had een paar loten gekregen van een vriend van me, die er meerdere had gekocht. |
| | | | Ik kocht ze zelf nooit, omdat ik niet geloofde dat ik ooit echt een kans zou maken. |
| | | | Ook nu had ik totaal niet verwacht dat ik iets zou winnen, en al helemaal niet veel. |
| | | | Toch ging ik even kijken op de website of mijn lot ook in de prijzen was gevallen. |
| | | | En toen stond mijn nummer daar onder de grootste prijs die er weggegeven werd. |
| | | | Ik heb het lotnummer wel tien keer nagekeken, want ik geloofde niet dat het de mijne was. |
| | | | Pas toen ik het lot ging inwisselen, en ze me feliciteerden, geloofde ik het. |
| | | | Ze gaven me champagne en bloemen, ik voelde me heel bijzonder. |
| | | Diploma/promotie | Op mijn 22e heb ik mijn rijbewijs in één keer gehaald. |
| | | | Ik had bijna een jaar les gehad, en vond het aan het begin best wel spannend. |
| | | | Nu was dan de dag aangebroken dat ik mezelf moest bewijzen. |
| | | | Ik was nog nooit zo zenuwachtig geweest, en daardoor vond ik het moeilijk goed te laten zien wat ik kon. |
| | | | De examinator zei weinig tijdens het examen, waardoor ik het gevoel kreeg dat ik het niet goed deed. |
| | | | Toen ik al vrij snel een foutje maakte zakte de moed me helemaal in de schoenen. |
| | | | Omdat ik dacht dat ik al gezakt was, was ik een stuk minder zenuwachtig. |
| | | | Blijkbaar heb ik daarna toch nog goed gereden, want toen we klaar waren hoorde ik dat ik geslaagd was. |
| | | | Ik kon bijna niet kon geloven dat het me echt gelukt was. |
| | | | Vorige week heb ik de promotie gekregen die ik graag wilde. |
| | | | Ik werk nu drie jaar bij dit bedrijf, en wilde heel graag hogerop komen. |
| | | | Ik had een afspraak met mijn baas, en wist niet waarover, dus ik was wat zenuwachtig. |
| | | | Ze vertelde me gelukkig dat ze de laatste tijd erg blij met me was. |
| | | | Ze begon over een vacature die ik voorbij had zien komen, en dacht dat die te hoog gegrepen was. |
| | | | Ze vond dat ik had laten zien dat ik daar goed in zou passen, en wilde me de functie graag aanbieden. |
| | | | Ik schrok er bijna van, omdat ik het helemaal niet aan had zien komen. |
| | | | Ik hoefde er niet eens over na te denken, ik heb meteen ja gezegd. |
| | | | De rest van de dag kon ik het eigenlijk nog steeds niet geloven. |



| | | | |
|---|---|---|---|
| Negatief | Verdrietig | Verlies familielid | Een paar maanden geleden is mijn zus plotseling overleden. |
| | | | Ze was op reis, en ze was daar in een groot busongeluk terecht gekomen. |
| | | | Toen ik het te horen kreeg van mijn vader, voelde het alsof mijn wereld instortte. |
| | | | Ik was thuis, en ik was me aan het klaarmaken om naar mijn werk te gaan, toen de telefoon ging. |
| | | | Mijn vader huilde aan de telefoon, en ik weet nog hoe ik meteen een steen in mijn maag voelde. |
| | | | Ik was erg close met mijn zus, en miste haar al erg terwijl ze op reis was. |
| | | | De gedachte dat ze nooit meer thuis zou komen was bijna onverdraaglijk. |
| | | | Het was misschien nog wel het ergste om het verdriet van mijn vader te voelen. |
| | | | Elke keer als de telefoon gaat word ik weer herinnerd aan dat moment. |
| | | | Mijn moeder is vorige maand overleden, na een jaar kanker te hebben gehad. |
| | | | Uiteindelijk belde mijn broer me om te vertellen dat ze was overleden. |
| | | | Die dag was één grote mix van emoties, het was echt heel heftig. |
| | | | Mijn vader, broer en ik hebben die dag veel over haar gepraat, en gehuild. |
| | | | Het was heel verdrietig om te realiseren dat ze er nu echt niet meer was. |
| | | | Ergens was er ook wel opluchting, omdat die zware periode achter de rug was. |
| | | | We waren ontzettend dankbaar dat we in haar laatste jaar nog zo veel leuke dingen hebben gedaan. |
| | | | Het was fijn om met mijn vader en broer over haar te kunnen praten, en ons verdriet te delen. |
| | | | Ik mis mijn moeder nog elke dag ontzettend. |
| | | Huisschade | Toen ik 's avonds laat thuis kwam, zag ik dat er ingebroken was in mijn huis. |
| | | | De deur stond open, en alles in het huis lag overhoop. |
| | | | Veel dure spullen waren weg, zoals de tv en de laptop. |
| | | | Ook waren de autosleutels weg, en was de auto meegenomen. |
| | | | Ik heb meteen de politie gebeld, die kwamen langs om wat onderzoek te doen. |
| | | | Ze konden niet meteen sporen vinden, en gingen weer weg. |
| | | | Ik voelde me heel machteloos, en onveilig in mijn eigen huis. |
| | | | Ik heb een vriend gebeld die die avond is blijven slapen, zodat ik niet alleen hoefde te zijn. |
| | | | De dader is helaas nooit gevonden, en mijn spullen heb ik nooit teruggekregen. |
| | | | Twee jaar geleden ben ik veel kostbaarheden verloren in een huisbrand. |
| | | | De brand was ontstaan bij de buren, en overgeslagen naar mijn huis. |
| | | | Het was avond, ik was alleen thuis, en ik zat op de bank TV te kijken. |
| | | | Ik herinner me nog dat het begon te stinken naar rook. |
| | | | Niet veel later ging het brandalarm af, en moest ik mijn huis ontvluchten. |
| | | | Ik kon niks meenemen, ik moest alles achterlaten. |
| | | | De brandweer kwam snel, en heeft de brand kunnen blussen. |
| | | | Helaas zijn mijn belangrijkste spullen, zoals mijn kinderfoto's en mijn oma's sieraden, verloren gegaan. |
| | | | Ik ben blij dat ik het overleefd heb, maar het blijft een groot gemis. |
| | Angst | Ongeluk | Een half jaar geleden ben ik ernstig gewond geraakt in een verkeersongeluk. |
| | | | Het was ochtend, het regende, en ik zat op de fiets onderweg naar mijn werk. |
| | | | Op een kruising heeft een auto die door rood reed me geschept. |
| | | | Ik kan me nog goed herinneren hoe de auto eruit zag: het was een blauwe Opel. |
| | | | Het ging heel erg snel, ik realiseerde het me pas toen ik op de grond lag. |
| | | | Er stond meteen een hele groep mensen om me heen die heel bezorgd keken. |
| | | | Ik voelde eerst even helemaal niks, maar na een paar minuten kwam de pijn. |
| | | | Ik kon niet meer praten of bewegen, ik was totaal in shock. |
| | | | Uiteindelijk heeft een ambulance me meegenomen naar het ziekenhuis. |
| | | | Ik ben van de trap gevallen en kwam heel verkeerd terecht. |
| | | | Ik was onderweg naar beneden met mijn handen vol schone was. |
| | | | Daardoor zag ik niet dat er iets uit mijn handen viel waar ik over struikelde. |
| | | | Van bijna helemaal boven rolde ik naar beneden, waar ik op de grond terecht kwam. |
| | | | Ik voelde meteen dat het niet goed was, en durfde me niet te bewegen. |
| | | | Daardoor kon ik geen ambulance bellen, en moest ik wachten tot er iemand thuis kwam. |
| | | | Ik weet niet hoe lang ik daar heb gelegen, het voelde als uren. |
| | | | Het deed overal pijn, en het werd steeds erger. |
| | | | Gelukkig kwam eindelijk mijn moeder binnen, die me heeft kunnen helpen. |
| | | Ziekte | Door een ernstige infectie ben ik bijna overleden. |
| | | | Ik had een simpele operatie gehad, die goed gegaan leek te zijn. |
| | | | Een week later werd ik verschrikkelijk ziek. |
| | | | Ik ben meteen naar het ziekenhuis gegaan, waar bleek dat ik een ernstige bacteriële infectie had. |
| | | | Binnen een dag lag ik op de intensive care en kon ik bijna niets meer. |
| | | | De artsen vertelden me en mijn familie dat er een grote kans was dat ik het niet ging redden. |
| | | | Het was alsof ik in een nachtmerrie zat, ik was zo vreselijk bang. |
| | | | Mijn familie kwam allemaal langs om afscheid te nemen. |
| | | | Gelukkig heb ik het net aan overleefd, al voel ik me een jaar later nog steeds niet de oude. |
| | | | Ik heb vorige maand te horen gekregen dat ik ernstig ziek ben. |
| | | | Ik voelde me al maanden niet goed en had veel tests gedaan. |
| | | | Van de dag dat de dokter me vertelde dat het niet goed was, weet ik alleen nog vlagen. |
| | | | Het kwam zo hard bij me binnen, dat ik niet veel heb onthouden. |
| | | | Gelukkig was mijn vriendin bij me, die heeft onthouden wat me verteld werd. |
| | | | Er zijn een aantal behandelingen die de symptomen kunnen verlichten. |
| | | | Helaas zijn er nog geen behandelingen mogelijk die de ziekte kunnen genezen. |
| | | | Daarom zal ik chronisch last van deze ziekte blijven houden en kom ik mogelijk in een rolstoel. |
| | | | Het lukt me niet goed om deze informatie te verwerken. |

Social Sciences

# The Political Strategy of Prawo i Sprawiedliwość

Christina Keßler

*Supervisor*
Dr. Carlos Reijnen (UvA)
*Reader*
Dr. Christian Noack (UvA)

## Abstract

This capstone analyses the political strategy and party behaviour of Prawo i Sprawiedliwość (PiS), the current governing party of Poland. It aims to explain how PiS managed to gain electoral support in the parliamentary elections of 2015 and how PiS continues to appeal to the public despite various acts which are deemed highly controversial. This work draws from scholarly theory concerning both the topics of political strategies of parties and the ongoing process of so-called democratic backsliding in Poland. Inspired by Strömbäck's (2007) four arena model, this capstone is divided into two sections. In the first part, this research project will use textual analysis of a press conference given by leading PiS politicians in September 2017 to illustrate the narrative the party tries to convey to its electorate. The second part analyses concrete acts of the party in order to shed light on PiS's political strategy. The findings demonstrate that PiS successfully employs a double strategy of reaching out to new voters by softening their image whilst at the same time satisfying its more radical core electorate. Furthermore, findings show that the party makes use of pre-existing narratives concerning 'political issues of the day', most notably during the 2015 refugee crisis.

Keywords and phrases: *Poland, illiberalism, party strategy, democratic backsliding, PiS*

# I. Introduction

Scholars are of the opinion that over the last few years, countries in Central and Eastern Europe (CEE), most notably Poland and Hungary, have been transforming into so-called 'illiberal democracies' (Cianetti et al., 2018; Rupnik, 2018). Research seeking to explain this phenomenon has largely focused on structural explanations, such as the larger historical and sociological context in which this process, oftentimes referred to as 'democratic backsliding', occurs (Rupnik, 2018). This research project takes a different approach, focusing not on structural factors, but rather aiming to gain a comprehensive understanding of the political strategy employed by the Polish political party Prawo i Sprawiedliwość (Law and Justice; PiS). PiS, which has been governing Poland since 2015, is the driving force behind the country's transition into an illiberal democracy (Cianetti et al., 2018; Grzymala-Busse, 2018; Rupnik, 2018). It is therefore questionable how PiS continues to appeal to the public despite a variety of acts that undermine institutions of liberal democracy and are perceived as highly controversial.

A prime example of such actions is the PiS party's restrictions of the Constitutional Tribunal, Poland's constitutional court. After taking office in November 2015, the new PiS government annulled the appointment of five judges to the Tribunal who had been nominated by the previous parliament (Szczerbiak, 2017). The Constitutional Tribunal subsequently ruled that two of the judges had indeed been nominated illegally, but also clearly stated that three out of the five appointments were constitutional (Szczerbiak, 2017). Nevertheless, the new government swore in five different judges who had been nominated by the new parliament (Szczerbiak, 2017). However, the then-President of the Constitutional Tribunal, Andrzej Rzepliński, only allowed two of the new judges to assume their duties (Szczerbiak, 2017). Trying to find a way around this impasse, the PiS government passed a law which both increased the number of judges required to make rulings and made it obligatory for the Tribunal to consider cases not in order of discretion but in the order in which those cases have been received (Szczerbiak, 2017). When the Constitutional Tribunal declared this law unconstitutional, the PiS government refused to publish the judgement in the official journal, which is a necessary step for rulings to become binding, and declared that the Tribunal had no constitutional power to review the law (Szczerbiak, 2017). This drew sharp criticism from various actors, such as the Venice Commission[1] which concluded that the actions of the PiS government endangered the rule of law as well as the democratic system of Poland (Rupnik, 2018; Szczerbiak, 2017).

When analysing the ongoing events in Poland, many scholars use the aforementioned term democratic backsliding (Cianetti et al., 2018; Przybylski, 2018). This phenomenon is not unique to Poland - indeed, various academics are of the opinion that we are currently witnessing an erosion of democracy all over the Central and Eastern Europe region

[1] The Venice Commission is an advisory body of the Council of Europe that deals with matters of constitutional law.

(Cianetti et al., 2018).

The goal of this research project is to examine the situation of Poland in the context of democratic backsliding, by focusing on the political strategy employed by PiS. Specifically, this research aims to answer the question of how PiS managed to gain electoral support and what strategies it employs to keep this support steady. Most importantly, this capstone questions the kind of narrative PiS aims to create about its role in contemporary Poland. Further questions address the concrete actions of the party, such as policies implemented by PiS. This capstone contributes to the academic discussion concerning rising illiberalism in the Central and Eastern European region and offers a new perspective by focusing less on structural explanations and more on contemporary political party strategies. Apart from its academic pertinence, the implications of this research are of societal and political relevance as they contribute to our conception of ongoing political processes in Poland by furthering our understanding of the success of PiS. Moreover, while the rise of illiberal democracy is not a distinctly European phenomenon, the research of its European instances can contribute to a general understanding of this phenomenon.

## II. Literature Review

The notion of democratic backsliding and what it constitutes is a question of ongoing debate (Waldner & Lust, 2018). The term implies incremental instead of radical change; “a process related to yet still distinct from reversion to autocracy” (Waldner & Lust, 2018, p. 94). According to Waldner and Lust (2018), backsliding always refers to a deterioration of some sort, while democratic backsliding specifically is defined as “a decline in the quality of democracy” (p. 94). Over recent years, the scholarly consensus seems to support the notion that we are currently witnessing a process of democratic backsliding in the CEE region (Cianetti et al., 2018; Grzymala-Busse, 2018; Rupnik, 2018). Cianetti et al. (2018), however, problematized this view, bringing forth various points of criticism, key among them being that calling the currently ongoing process ‘democratic backsliding’ implies a preceding successful democratisation. While the concept of democratic backsliding in CEE remains controversial, the notion that contemporary Poland and Hungary can be seen as illiberal democracies remains largely undisputed and is even embraced by leading political figures such as Hungarian Prime Minister Victor Orbán, who publicly declared himself in favour of an ‘illiberal state’ (Rupnik, 2018). The concept of illiberal democracy can be defined as a regime which falls somewhere between democratic and nondemocratic, and which combines both democratic and illiberal institutions and practices (O’Neil, 2015). In illiberal regimes, the rule of law might be in place, but institutions resting on the rule of law (as well as the rule of law itself) are weakly institutionalized and not well-respected (O’Neil, 2015). The Polish turn towards illiberalism is a very recent development and represents one of the most extreme cases of deteriorating democracy in the region (Cianetti et al., 2018). In order to fully comprehend this, one must be familiar with the historical background of the country and of Central and Eastern Europe.

Poland is the largest state in the region and is considered to be especially important from a geopolitical perspective (Cianetti et al., 2018). Formerly partitioned by Germany, Austria, and Russia, Poland was occupied by Nazi Germany and Soviet Russia during the Second World War and belonged to the Soviet Bloc until 1989 (Mucha, 2006). The agricultural country has been linked to Eastern Europe through its internal social structure, its economy, and its politics, while it is connected to Western Europe by its Latin Christianity, its alphabet, and its cultural aspirations (Mucha, 2006). Poland proved more resistant to communism than other countries in the region, a fact that has been attributed to its culture, which Rupnik calls a “combination of nationalism and Catholicism” (Rupnik, 2018, p. 28). The country’s democratic transition after the collapse of the communist regime in 1989 proved a challenging task (Przybylski, 2018). Inflation rose to 640 percent and the newly introduced reforms left many people losing their wealth, their social status, and their trust in the new institutional arrangements (Kowalewski & Rybinski, 2011; Przybylski, 2018). In the following years, Poland underwent ambitious reforms (Przybylski, 2018), and was the first among the countries in the CEE region to recover from the sharp economic decline following the transition of 1989 (Kowalewski & Rybinski, 2011), experiencing a period of continuous economic growth from 1992 onwards (Kowalewski & Rybinski, 2011; Mucha, 2006; Przybylski, 2018).

In 1997, Poland became a member of NATO and the country joined the European Union in 2004 (Mucha, 2006; Przybylski, 2018).

Due to Poland's economic development as well as the fact that it was the only EU member state to survive the 2008 financial crisis without a recession, Poland has been heralded as an example for successful post-communist democratic reforms (Przybylski, 2018; Rupnik, 2018). However, the recent changes (both legislative and institutional) brought about by PiS have evoked fears that the accomplishments of the successful democratic reform might be reversed (Przybylski, 2018). Poland's path to illiberalism and the role of PiS in this process is traced by Anna Grzymala-Busse (2018), who claims that contemporary Poland must be defined as an illiberal democracy, because it remains a democracy in which the critical institutions of liberal democracy such as the courts and the public media are no longer able to fulfil their purpose. This argument aligns with the definition of illiberal regimes as outlined by O'Neil (2015). Seeking to understand the surge of illiberalism in the CEE region, Rupnik (2018) points out the decoupling of liberalism and its focus on individual freedom from democracy. While democracy has historical antecedents in Central and Eastern Europe, liberalism is understood as a foreign concept imported to the region after 1989 (Rupnik, 2018). Rupnik (2018) argues that Poland and Hungary view liberalism as a model of society which is not only imposed but also failing in critical situations like the 2015 refugee crisis. Additionally, he points towards the loss of confidence in the institutions of liberal democracy and the fears of identity loss in Polish society (Rupnik, 2018). In the context of rising illiberalism, Przybylski (2018) introduces different national narratives. Those include a narrative of sovereign democracy (according to which democratic legitimacy places the actions of a government above criticism) and the 'antemurale myth' (according to which refugees and migrants mobilize on a march through Europe and must be stopped by countries such as Poland). Fomina & Kucharczyk (2016) further point towards Norris' theory of "contemporary authoritarian populism" (p. 58) according to which we can understand the current developments in Poland as a backlash against long-term, ongoing social change. They also name various PiS campaign strategies as well as circumstantial factors which were present during the electoral campaign 2015, such as weakness of the opposition, the 'Poland in ruins' narrative employed by PiS, the political strategy of softening the party's image, campaign promises, and a focus on identity politics during a time in which the topic of migration dominated public discourse (Fomina & Kucharczyk, 2016). In combination, these factors contributed to the electoral success of PiS in the parliamentary election in 2015.

Party behaviour and electoral strategy have long been a focus of scholarly research (Brysk, 1995; De Vries & Hobolt, 2012; De Sio et al., 2018; Somer-Topcu, 2015). While the research presented in the following paragraphs is not directly linked to the Polish context, it provides significant information about political strategy on a general level. In this research, these theories on political strategy will be synthesized and applied to the Polish context in order to further understand the party behaviour of PiS. When thinking of a political strategy, we oftentimes think first and foremost of campaigning and elections. While elections are an important part of party competition, they are by no means the only one. Strömbäck (2007) uses the 'four arena model' in order to conceptualize the different areas in which political parties compete: an *electoral arena*, a *parliamentary arena*, a *media arena*, and an *internal arena* (Strömbäck, 2007). Depending on the larger context in which they are competing, political parties adopt different strategies for each of the respective areas. (Strömbäck, 2007). A strategy commonly employed by political parties in the *electoral arena* is elaborated on by Somer-Topcu (2015): many parties try to appeal to broad audiences in order to gain more votes. Political parties can achieve this goal in a variety of ways, for instance by taking distinct positions on various issues from different sides of the political spectrum; by using candidates with particular profiles to strengthen or soften their positions; or by beclouding their own policy positions (Somer-Topcu, 2015). Such strategies help parties win votes (Somer-Topcu, 2015). De Vries and Hobolt (2012) link the study of political strategies to theories of issue evolution which emphasize the importance of a changing issue-agenda (2012). 'Political losers' can advance their position in a party system by taking up an issue entrepreneurial strategy, thus introducing new issue dimensions (De Vries & Hobolt, 2012, p. 246). Research shows that voters are responsive to such strategies and reward issue entrepreneurs (De Vries & Hobolt, 2012). The importance of specific issues is also the focus of the research of De Sio et al. (2018), which introduces the issue yield model, according to which political parties select issues based on the consideration of whether they are positively associated with the party and whether a stance is widely shared by the electorate. Reeves et al. (2006) highlight the importance of an increasing focus on the demands of the electorate: they establish that British political parties have become more marketing-oriented and that the importance of ideology has declined. This, however, results in tension between making responsible long-term decisions and being exclusively voter-driven (Reeves et al., 2006). Johnston et al. (2018) also focus their research on political parties in Britain and conclude that voters often base their decisions about a political party on their feelings towards a party leader instead of specific policies. If party leaders change, voters might very well change how they feel about a party (Johnston et al., 2018). While both the research of Reeves et al. (2006) and Johnston et al. (2018) focus on a British context, they can give us indicators on how political parties might behave in general.

However, changes in an electorate's perception of a political party cannot always be attributed to concrete policies or changes in the party structure. Brysk (1995), in contrast to the previous theorists, highlights the importance of stories and the role of narrative in politics. In order to understand political change, we must comprehend that the decision of the electorate is not only based on fixed needs but that the adoption of new narratives concerning interests and identities can lead to paradigm changes which influence voting behaviour (Brysk, 1995). While the term narrative is oftentimes associated with fiction rather than with political science, it has over time become an invaluable tool in social science methodology (Patterson & Monroe, 1998). In the field of political science, the narrative plays an essential role in one's perception of one's political reality and hence greatly influences political behaviour (Patterson & Monroe, 1998). Theories of narrative construction can, therefore, be understood as yet another set of theories further explaining political strategy and party behaviour: the narrative PiS constructs and conveys to its electorate is an essential part of its political strategy.

## III. Methodology

This interdisciplinary capstone combines the fields of political sciences and communication studies in order to gain a comprehensive understanding of the political strategy employed by PiS. In order to contextualize the political behaviour of PiS, scholarly literature on the current political situation and the rise of illiberalism in Poland is examined. In addition to this, this capstone draws from scholarly theory concerning political strategy and party behaviour. Its structure is inspired by Strömbäck's (2007) four arena model. According to this model, political parties acting in multiparty systems compete in four areas: the *electoral* arena, the *media* arena, the *internal* arena, and the *parliamentary* arena. Analyzing merely the electoral arena is limiting and does not adequately reflect the aims of a political party, as a party's aims go beyond electoral victory. Each arena has respective strategic goals, primary actors and decision concerns associated with it (Strömbäck, 2007). It is important to note that the strategies pursued in different arenas can very well contradict each other: the positions a party takes on the campaign trail are not always reflected in the parliamentary behaviour of that very same party (Strömbäck, 2007). It would also be misguided to conceptualize the four arenas as strictly separate from each other; in practice, the arenas overlap. Strategic decisions regarding a party will often affect more than one of the arenas, and the arenas influence each other as well (Strömbäck, 2007). Keeping in mind that the arenas are not isolated entities but rather deeply intertwined, the four arena model can serve as a starting point to gain a comprehensive view of party behaviour. Inspired by the four arena model, this capstone is divided into two sections: one examining the narrative employed by PiS, the other analyzing the party's concrete actions.

The first section (Narrative) is influenced by Strömbäck's conception of the *electoral* and the *media arena*. In those arenas, it is of the utmost importance what kind of story a political party tells to its electorate. In order to gain a comprehensive understanding of this story, the following research makes use of content analysis. The narrative was derived from a textual analysis of an official press release by PiS which covers a joint press conference by the then Prime Minister Beata Szydło and PiS party leader Jarosław Kaczyński in September 2017 (Chancellery of the Prime Minister, 2017). This press conference was chosen as it was given by the then most important political leaders of PiS, and clearly frames the policies of the party. In order to determine the narrative from the official press release, an analysis guided by emergent codes was carried out (Chancellery of the Prime Minister, 2017). Afterwards, the content of the press release was organized in a matrix in which every paragraph of the text was thematically sorted into at least one of the categories that the coding process had brought forth (Appendix A). Then, the findings for each of the emergent codes were gathered and used to describe how a certain topic area was framed in the press conference. Subsequently, an analysis was carried out that examined which topics had been emphasized and which topics had been barely mentioned. The analysis also put the findings in context with current scholarly literature concerning the party strategy of PiS.

The second section (Action) is inspired by Strömbäck's conception of the *parliamentary* and the *internal arena*. Through an analysis of concrete policies and actions of PiS, this research seeks a holistic understanding of PiS's political strategy, and limits itself to three concrete actions, prioritizing those which are widely regarded as the most prominent by current literature on the PiS government. Concretely, these actions are social spending pledges made by the party, decisions taken regarding leadership positions in the current Polish government, and PiS's migration policy (Fomina & Kucharczyk, 2016; Przybylski, 2018). Those actions are thoroughly described and contextualized in a findings section, followed by an analysis of the actions supported by scholarly literature regarding political strategy and party behaviour.

## IV. Narrative

When trying to understand political outcomes the assumption of a rational choice model, in which actors base their decision on fixed needs, falls short (Brysk, 1995). Rather, one must take the importance of ideas and normative beliefs into consideration, because they have the power to shape agendas, build identities and shift paradigms (Brysk, 1995). According to Brysk (1995), "interests are not fixed needs, but rather deeply subsumed stories about needs, [...] [and] symbolically mobilized political actors can create new political opportunities by revealing, challenging, and changing narratives about interests and identities" (p. 561). Therefore, the ability to create a compelling narrative or to insert oneself in an already existing one an essential part of bringing about political and collective change and influences the success of political actors (Brysk, 1995). The term narrative can be defined as a story describing the way in which one constructs facts into a coherent manner in order to make sense of reality (Patterson & Monroe, 1998). When looking at Strömbäck's four arena model, narrative is most closely related to the electoral and the media arena. Strömbäck (2007) distinguishes different strategic goals for the different arenas in which political parties compete: in the *electoral arena*, the goal is "to maximise electoral and voter support"; in the *media arena*, the goal of a political party is "to maximise positive publicity" (p. 59). Both of these strategic goals require that the political party engage in an act of storytelling by conveying a new, compelling narrative to the public. As Brysk (1995) points out: "We think about politics in stories, and our consciousness is changed when new stories persuade us to adopt a new paradigm" (p. 561). Subsequently, understanding the narrative PiS tries to convey to the Polish electorate is an essential part of analysing the party's political strategy. Narrative, as understood in this research project, is based on Somer and Gibson's definition of narrativity (Patterson & Monroe, 1998). Somer and Gibson's definition, which is particularly relevant to social sciences research, contains four features (Patterson & Monroe, 1998): *1) relationality of parts* (events need to be placed in relation to other events), *2) causal emplotment* (events in the narrative have a causal relationship with each other), *3) selective appropriation* (some elements are incorporated while others are omitted), and, taken together as a fourth feature, *4) temporality, sequence*, and *place* (referring to the way in which the elements are located vis-à-vis each other) (Patterson & Monroe, 1998). Based on this definition, this section will analyze the narrative brought forth by PiS and explain how it contributes to the party's popularity. Firstly, the narrative will be derived through the analysis of an official press release covering a joint press conference in November 2017 by then-Prime Minister Beata Szydło and the leader of PiS, Jarosław Kaczyński. After the analysis of the aforementioned primary source, the findings will be put into context with scholarly literature concerning PiS. Finally, the chapter will answer the question of how the narrative employed by the party contributes to its appeal.

In order to contextualize the press conference, some information on the political setting in which it takes place is needed. While the current PiS government is the first instance of a single-party majority government in post-communist Poland's history, PiS has previously been part of the Polish government from 2005 to 2007 (Jaskiernia, 2017). However, the coalition led by PiS fell apart in 2007, triggering early elections (Fomina & Kucharczyk, 2016). During these, a centrist government under then Prime Minister Donald Tusk and his party Platforma Obywatelska (Civic Platform; PO) was elected and subsequently "moved Poland back toward the mainstream of European Politics" (Fomina & Kucharczyk, 2016, p. 59). Between 2006 and 2014, the Polish electorate chose PO over PiS eight consecutive times in elections at the local, national, and European level (Fomina & Kucharczyk, 2016; Jaskiernia, 2017). As Fomina and Kucharczyk (2016) describe, an electoral victory of PiS had been deemed increasingly unlikely by experts: "Analysts had begun to regard PiS as unelectable, and dismissed its authoritarian longings and conservative social ideology as lacking appeal outside older, less educated, and poorer sections of Polish society" (p. 59). However, in 2015, PiS managed to win the Sejm (the lower house of Polish parliament), the Senate and the presidency, and became the first electoral victor in the country's democratic history to form a government without needing coalition partners (Fomina & Kucharczyk, 2016; Jaskiernia, 2017). While many political scientists consider PiS to be a 'far-right' party, delineating the party with a clear ideological label remains a challenge (Jaskiernia, 2017). Positions of the party entail an increase in social spending, increased taxation of the wealthy, and the re-nationalization of key economic sectors (Jaskiernia, 2017). Kaczyński, the party's current leader, stated that PiS is opposed to "immigrants, gays, feminists, liberals, and most foreigners" (Jaskiernia, 2017, p. 238). Further, PiS makes a harsh distinction between who does and does not belong to the Polish people and explicitly opposes ideas of multiculturalism (Fomina & Kucharczyk, 2016; Jaskiernia, 2017). The press conference analyzed in the following section is set in November 2017, two years after PiS's electoral victory in 2015. It is an ideal source from which to derive PiS' narrative considering that it was given at the occasion of the two-year hallmark of PiS government rule, and frames the actions of the party during the past two years. The press conference was also held by two of the most prominent PiS politicians at the time, respectively occupying two of the most important positions: the head of the party (Kaczyński) and the head of government (Szydło).

## V. Joint press conference by Szydło and Kaczyński

The analysis of the press conference was guided by emergent codes. The codes were the following: development of the Polish economy, reclaiming the country and feeling at home in Poland, outside pressures, ordinary people and families, equal opportunities and dignified life, failures of the past, judiciary reform, security, and success. The content of the press release (Chancellery of the Prime Minister, 2017) was analyzed using a matrix which thematically sorted every part of the press release into at least one of the aforementioned categories (Appendix A). The following analysis clarifies how the press conference frames the different topic areas mentioned above. Hence, this analysis does not only show *which topics* PiS party leaders choose to talk about, but it also analyses *how* these different issues are framed by the party.

### Findings

#### Development of the Polish economy.

The press conference spends large parts focusing on the development of the Polish economy. Prime Minister Beata Szydło, for example, emphasizes that strengthening the Polish economy is one of the primary goals of the PiS government, stating: "Family, development and security are the three pillars of our governance" (Chancellery of the Prime Minister, 2017). Various PiS government initiatives with economic dimensions are mentioned, such as lowering the retirement age and "the flagship PiS government project, namely Family 500+ programme" (Chancellery of the Prime Minister, 2017), which provides child benefits to Polish citizens. It is emphasized that "Putting Polish economy back on a sound footing" (Chancellery of the Prime Minister, 2017) is one of the main priorities of the PiS government. Regarding the state of the economy in previous years, Beata Szydło states that : "PiS is freeing Polish economy and Polish entrepreneurs from the shackles of inability, bureaucracy, all that had blocked the energy of Polish economy" (Chancellery of the Prime Minister, 2017).

### Equal opportunities and dignified life.

The notion of bringing about "equal opportunities" and a "dignified life" (Chancellery of the Prime Minister, 2017) for the Polish people is mentioned several times in the press release, albeit not as often as, for example, the development of the Polish economy. According to the text, "The fundamental assumption of the programme proposed by the Law and Justice was to ensure that all Poles - regardless of their place of residence or their occupation - have equal opportunities" (Chancellery of the Prime Minister, 2017). Notably, the press release does not only mention the goal of giving people a life of dignity but explicitly uses the expression of "restoring" a life of dignity twice, therefore implying that over the previous years, Polish citizens had been deprived of the very dignity the PiS government is now giving back to them (Chancellery of the Prime Minister, 2017).

### Failures of the past.

While the press release gives few concrete examples of past failures, it generally conveys the impression that the living conditions for Polish citizens have been improving under PiS rule. There is one example of past failures which is further elaborated: it is mentioned that PiS is pursuing policies which make it possible for "a majority of Polish families, which earlier had not benefited from economic development in Poland" (Chancellery of the Prime Minister, 2017), to finally do so. In the press release, this issue is directly related to corruption and fraud by Jarosław Kaczyński:

> *"He then pointed out that two years ago, representatives of PiS had found out that the level of all kinds of corruption and fraud in Poland is high. We had concluded that if we harness, even partially, these pathologies, we would be able to implement social policy that would be beneficial for groups that had benefited little, if at all, from changes taking place in Poland. This diagnosis has turned out to be true"* (Chancellery of the Prime Minister, 2017)

### Judiciary reform.

While the judicial reform sought by the PiS government is mentioned, the idea is not developed at length in the press release. The controversy surrounding judicial reform is not brought up at all, instead, the reform of the judiciary wing is presented as something positive:

> *"I hope very much that we will complete the process of judiciary reform, - Prime Minister Beata Szydło declared. She highlighted that the judiciary reform makes no sense unless it is radical and gives Polish people the feeling that courts are given back to them and that citizens are treated fairly"* (Chancellery of the Prime Minister, 2017).

By framing completing the reform of the judiciary wing as something necessary for Polish citizens, PiS is able to relate their agenda directly to a so-called "giving back" of the courts and fair treatment (Chancellery of the Prime Minister, 2017).

### Ordinary people and families.

No other topic is as frequently mentioned as the topic of ordinary Polish people and families; PiS presents itself as a representative of the "ordinary people" (Chancellery of the Prime Minister, 2017), a term which is used repeatedly in the press release.. The press release emphasizes that the PiS programme is directly built on the needs of "ordinary people" (Chancellery of the Prime Minister, 2017) and was created in dialogue with them. Every policy area is directly related to improved living conditions for ordinary people and Polish families, who benefit from programmes such as 500+. Repeatedly, a notion of restoring dignified living conditions for ordinary Poles and Polish families is brought up. When Prime Minister Szydło states that, "We wish to cordially thank all Polish people who allow for the good change to happen and who help us to put it in effect" (Chancellery of the Prime Minister, 2017), she notably explicitly talks to the Polish people and expresses her gratitude for their support.

#### Outside pressures.

Outside pressures on the PiS government are only mentioned twice throughout the entire press conference. The first time, they are merely hinted at, when Kaczyński thanks Prime Minister Szydło for "never failing us in difficult moments, both at the beginning of her term and later, despite the different pressures put on her" (Chancellery of the Prime Minister, 2017). Later, the press release gets more specific and mentions the dispute over migration with the EU, stating that: "We have won the migration dispute with the European Union", and effectively framing the situation as a victory for Poland and Polish influence over the discourse (Chancellery of the Prime Minister, 2017).

#### Reclaiming the country and feeling at home in Poland.

Repeatedly the press release expresses a notion of 'reclaiming' the country. This notion is connected to several policy fields, such as economic development, security issues, and judicial reform. What exactly Poland is being 'reclaimed' from remains unsaid. However, the notion of 'reclaiming' the country is connected to the living conditions of ordinary Polish people and families, for whom the country must be 'reclaimed'. The press release emphasizes that significant changes are needed in order to ensure that "Poland is safe and is growing, and all Polish families feel that they are rulers of their own country" (Chancellery of the Prime Minister, 2017), implying that this is not the case under the current circumstances.

#### Security.

Security, according to the press conference, is one of the three pillars of PiS governance (together with family and development), and is brought up in relation to three specific issues. One of them is the Warsaw NATO summit, which Beata Szydło described as a "breakthrough" (Chancellery of the Prime Minister, 2017). The second one is government support for the armed forces as well as the police forces, which are framed as having contributed to the "everyday sense of security of Polish people" (Chancellery of the Prime Minister, 2017). Thirdly, Beata Szydło notes that "Poland is perceived as a terrorism-free country" (Chancellery of the Prime Minister, 2017); this notion is directly followed by commenting on the migration policy of the EU and the former PO government, implying that Poland's status as a terrorism-free country is due to the harsh migration policy of the PiS government. However, in connection with the migration crisis, Beata Szydło also points out that Poland would be delivering humanitarian help.

#### Success.

The press conference frames the two years of PiS government as a success story. Several government projects are praised and are said to have fulfilled their purpose. Kaczyński goes as far as to claim "that all the undertakings of PiS government have been successful" (Chancellery of the Prime Minister, 2017). This framing makes sense, as the press conference took place two years after the PiS government started their work. The press conference conveys the notion that the PiS government is a successful administration, which keeps its promises and betters the living conditions of ordinary people. The work of the administration is by no means finished though: "As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely" (Chancellery of the Prime Minister, 2017). It is made clear that the "good change"(Chancellery of the Prime Minister, 2017), the slogan of PiS's 2015 election campaign, is not completed yet.

## Analysis

### Joint press conference by Szydło and Kaczyński

The matrix shows that three issues dominate the press release: ordinary people and families, the successes of the PiS government, and the development of the Polish economy. According to the press conference, these issues are closely related. The PiS government is presented as a government that puts improving the living conditions of the so-called "ordinary people" (Chancellery of the Prime Minister, 2017) at the front and centre of its programme. PiS's claim to only represent 'the real people' in contrast to an elite establishment is a textbook definition feature of populism (Grzymala-Busse, 2018). However, while PiS has denounced the Polish "post-1989 elite establishment" as a "[corrupt] cartel" (Grzymala-Busse, 2018, p. 96) in the past, this press release stays vague and does not specify from whom the country must be reclaimed.

Large parts of the press conference are devoted to the development of the Polish economy - more specifically focusing on various economic and social policies introduced by the PiS government, such as the 500+ child benefit programme and the lowering of the retirement age. This is insofar striking as it is expected, as it allows the PiS to join in a historic tradition: "From the very start of the transition era in Poland, political narratives and policy decisions have rested to a large extent on an economic rationale, with the promise of prosperity at its center" (Rupnik, 2018, p. 54). While the Polish economy has been performing well over the last few years (Przybylski, 2018; Rupnik, 2018), the press conference argues that the "ordinary people" (Chancellery of the Prime Minister, 2017) have not sufficiently benefited from this development, an issue that the PiS government now aims to tackle. In fact, the claim that "other parties have been hoarding the fruits of the transition, forsaking social solidarity with those who had reason to resent the new system" (Przybylski, 2018, p. 56) is by no means a novelty, but an argument Kaczyński has been making since the 1990s (Przybylski, 2018). The PiS government, led by Kaczyński, thus argues that its various initiatives have already benefited the Polish people economically and will continue to do so in the future.

The press conference does not tell a story of struggle, but a story of triumph. The two years of PiS governance are framed as a success story that is set to continue. Szydło and Kaczyński focus on the accomplishments of the PiS government, placing a strong emphasis on economic and social pledges the party has managed to fulfill. In order to derive the distinct frame which was chosen for this press conference, it is important to distinguish not only which political issues were emphasized, but also which issues were just briefly mentioned or entirely ignored in the telling of this success story. Two topics that provoked controversy both within Poland and abroad during the first two years of PiS governance were the state of the rule of law and the restrictions the government imposed on reproductive rights. The topic of judicial reform is only brought up twice in the press conference: once presented as a promise to the Polish people that must be fulfilled in the two years ahead, and once as a reform that "makes no sense unless it is radical and gives Polish people the feeling that courts are given back to them and that citizens are treated fairly" (Chancellery of the Prime Minister, 2017). The approach of PiS to the reform of the judiciary wing could be considered radical - shortly after being elected, the PiS government started attacking judicial independence and Poland's Constitutional Tribunal through legislation, while using procedural obstructions to stall the announcement of court vacancies (Przybylski, 2018; Rupnik, 2018; Fomina & Kucharczyk, 2016). These actions provoked nation-wide protests, within Poland most notably from a grassroots-initiative called Komitet Obrony Demokracji (Committee for the Defense of Democracy; KOD), lawyer's groups, and other Non-Governmental Organisations (Fomina & Kucharczyk, 2016; Grzymala-Busse, 2018). Internationally, the efforts of the PiS government drew sharp criticism from the European Union:

> *"Late in December 2015, Poland became the first EU member state to be a subject of the Union's new "pre-Article 7 procedure," adopted just the year before, for looking into possible breaches of EU standards, including harms to the rule of law. In January 2016, the European Commission launched an official probe."* (Fomina & Kucharczyk, 2016, p. 64).

While the press conference brushes over the topic of judicial reform rather quickly, it entirely fails to mention the controversy concerning the PiS government's restrictions on reproductive rights. A citizen's bill of March 2016 which would have enacted a full ban on abortion, but was ultimately not voted on in parliament after mass protests by Polish citizens (Fomina & Kucharczyk, 2016). To claim the problem was no longer relevant however would be incorrect, as "issues remain subjects of heated public controversy, with attendant demonstrations and petition drive" (Fomina & Kucharczyk, 2016, p. 65). This is also due to the Health Ministry's public announcement to consider restricting access to contraception (Fomina & Kucharczyk, 2016). Likely the two aforementioned topics were not elaborated on in the press conference due to the domestic criticism they provoked, which, in the case of both judicial reform and of abortion rights, led to mass protests by Polish citizens.

Apart from these two issues, there is another topic which strikingly is not elaborated on in the press conference: migration. Migration is only explicitly mentioned twice throughout the entire conference: first, when Szydło and Kaczyński emphasize that the PiS government will not take part in “the misguided EU migration policy” (Chancellery of the Prime Minister, 2017) while pointing towards the fact that terrorism does not occur in Poland; secondly, when they declare that the PiS government has “won the migration dispute with the European Union” (Chancellery of the Prime Minister, 2017) and changed the discourse concerning migration. While the issue of migration, therefore, is certainly mentioned, it is not discussed, especially compared to topics such as economic development. This is remarkable, as the party’s position on migration is generally perceived as one of PiS’s political assets (Narkowicz, 2018). When trying to understand the electoral victory of PiS in 2015, scholarly literature regularly points to the party’s position on migration as one of the reasons for its appeal, as its stance resonates with the Polish public (Fomina & Kucharczyk, 2016; Narkowicz, 2018; Przyblyski, 2018; Rupnik, 2018). Rupnik (2018) attributes the appeal of PiS largely to the fact that “Central European countries perceive the redistribution of migrants across national borders [...] as an attempt to impose on them a multicultural model of society that they consider a failure” (p. 33). According to Fomina and Kucharczyk (2016), “[Kaczyński’s] hard line on refugees, verging on xenophobia, won over people who normally would never have voted for PiS” (p. 62) in light of the 2015 migration crisis. The omission of an elaborate discussion on the topic of migration during the press conference is worthy of scrutiny, especially considering the importance of the issue for PiS (Narkowicz, 2018). There are other issues generally perceived as PiS’s core issues that were ignored during the press conference. According to Fomina and Kucharczyk (2016),

> *“Although Law and Justice was elected on a platform of generous socioeconomic promises, its dominant position on the righthand side of the political spectrum and its resilience despite years in opposition came from its strong stance on issues connected to national identity and sovereignty and from its bond with the Catholic Church.”* (p. 66).

However, the Church was not brought up even once during the press conference, nor were left-liberal values (such as secularism, feminism, LGBTQ rights, multiculturalism) towards which PiS often presents itself in opposition (Rupnik, 2018).

The fact that the press release fails to elaborate on issues which are generally perceived to be crucial to the party’s identity and of high importance to its core electorate, suggests PiS trying to appeal to new parts of the electorate - people who traditionally would not have voted for PiS and are not necessarily swayed by the ideology that PiS puts forth. Indeed, the narrative presented in this press conference emphasises not ideological, but rather socioeconomic considerations for supporting the PiS government. Controversial topics are largely avoided or brushed over. For a party which often verges on xenophobia (Fomina & Kucharczyk, 2016), the little stab that Szydło and Kaczyński take at EU migration policy appears almost negligible. The strategy of focusing not on overtly ideological issues but rather appealing to more moderate parts of the electorate has also proven successful for PiS in the past: the conscious deradicalization of the party image has been on the agenda of the party since the 2015 elections. Back then, “Running on the slogan “Good Change”, PiS leaders called for compassionate conservatism, and sought to offer undecided voters an alternative to the “boring” PO” (Fomina & Kucharczyk, 2016, p. 61-62). This strategy was rewarded by the electorate: in 2015, PiS “reaped the benefits of a more moderate appeal to the general public” (Przyblyski, 2018, p. 56). PiS’s strategy to deradicalize its message as much as possible in order to appeal to a broader public also makes sense regarding the setting of the press conference, considering that it was an official event with the then two most prominent PiS party politicians, and therefore likely to draw the attention of a broad audience.

Whilst PiS’s ideology is somewhat softened, radical features of the party’s politics keep shining through if one knows what to look out for. This becomes especially apparent in the repeated notion that the country needs to be ‘reclaimed’ and ‘restored’, insinuating that the country must have been at least somewhat destroyed by previous governments. Fomina and Kucharczyk (2016) do indeed attribute the success of PiS to the fact that the party managed to successfully convey the narrative of a “Poland in ruins” (p. 60) to the Polish electorate. According to this narrative - which ignores the economic progress Poland experienced after it had recovered from transition - Poland has to be restored and rebuilt “after the devastation allegedly wrought by PO’s eight-year rule, or even the quarter-century of Poland’s democratic transformation” (Fomina & Kucharczyk, 2016, p. 60). The success of this narrative can be partially credited to the reforms of the 1990s. The privatization of whole economies over a relatively short period of time, “often with a contempt for the rule of law and societal sentiments about equity” (Rupnik, 2018, p. 34), resulted in a general distrust of the political elite (Kowalewski & Rybinski, 2011; Przybylski, 2018). While deradicalization may take place concerning the party’s explicit ideology, PiS does not hide the fact that its goal is “to change Poland completely” (Chancellery of the Prime Minister, 2017).

The narrative presented here seamlessly fits the definition of narrativity by Somers and Gibson, which includes *relationality of parts, causal emplotment, selective appropriation*, and *temporality, sequence* and *place*. (Patterson & Monroe, 1998). The press conference places events in a causal relationship to other events and therefore meets the first two features: *relationality of parts* and *causal emplotment*. During the press conference, for example, Szydło and Kaczyński not only suggest that Poland is being reclaimed by ‘ordinary citizens’, but they also place this ‘reclamation’ in a causal relationship to the efforts of the PiS government. The third feature of the definition, *selective appropriation*, is likewise apparent in the conference. Issue areas, such as the controversy regarding abortion laws that have been omitted in the press conference, as explained above. The fourth requirement, the elements of *temporality, sequence*, and *place*, is also met. Exemplar for this is the story that previous (*temporality*) governments have brought Poland (*place*) in a position where it now needs to be reclaimed by its citizenship, a task that the current PiS government is now successfully taking on (*sequence*).

When examining the narrative presented in the press conference, one notices that the radical elements of PiS’s ideology, while hidden, are still noticeable for anyone familiar with the party’s history, suggesting a double strategy. The traditional ideological core of PiS with its authoritarian and xenophobic longings still very much exists, as certain comments and allusions in the press release demonstrate. However, the party aspires to send a message that does not only appeal to its traditional electorate and subsequently softens its message.

## VI. Action

Understanding the narrative a political party employs is an integral but ultimately insufficient step towards understanding its strategy. In order to draw a holistic picture of PiS’ political strategy, one must also examine the party’s concrete actions in and outside of parliament. This dimension is reflected in Strömbäck’s four arena model through the *parliamentary* and the *internal arena* (Strömbäck, 2007). According to the model, the goal of a political actor in the *parliamentary arena* is “to maximize parliamentary influence” (Strömbäck, 2007, p. 59), while the strategic goal in the *internal arena* is “to maximise internal cohesion” (Strömbäck, 2007, p. 59). These two arenas make apparent that, in addition to an analysis of PiS’s political message, an analysis of PiS’s concrete actions is necessary. Again, it must be noted that while narrative and concrete party actions are treated as two separate dimensions in this capstone, the two areas are intertwined, and overlap and influence each other in a variety of ways.

In the following section, three concrete actions will be examined: 1) social spending policies which have been implemented by PiS, 2) the decision to select Beata Szydło as Prime Minister instead of party leader Jarosław Kaczyński, and 3) the refusal to take in refugees and the following migration policy dispute with the European Union. These three actions have been selected as that they all occurred within the two years after the PiS government came into power in 2015, and are regarded to be cornerstones of PiS popularity with the Polish electorate (Fomina & Kucharczyk, 2016; Narkowicz, 2018; Przybylski, 2018; Rupnik, 2018). Each action will first be contextualized and elaborated on, answering the question of what exactly the PiS government’s actions in specific issue areas were. Then, the strategic benefit of each concrete action will be analysed in connection to the relevant scholarly literature. To conclude the chapter, how the aforementioned decisions by PiS contributed to the party's appeal will be summarized.

## VII. Concrete actions of the PiS government

### Findings

#### Social Spending.

When Beata Szydło ran for Prime Minister in 2015, her platform included generous social spending pledges (Fomina & Kucharczyk, 2016). The two most prominent examples of these are a child benefit programme and the promise to lower the retirement age after it had been raised in previous years (Fomina & Kucharczyk, 2016). Once elected, the government maintained the promises they had made in the election campaign. The programme 500+ was implemented, financially benefiting families with children (Grzymala-Busse, 2018; Szczerbiak, 2017). The unpopular 2011 decision by Donald Tusks' government to gradually raise the pension age to 67 was reversed as the age of retirement was moved to 60 for women and 65 for men (Grzymala-Busse, 2018; Szczerbiak, 2017). While both of these social spending pledges were rather expensive promises to fulfill, one must note that the Polish economy has been performing well recently, with unemployment rates dropping and investments increasing (Szczerbiak, 2017). As Anna Grzymala-Busse (2018) points out: "Claims from the opposition that such measures could ruin the public finances ring hollow against a background of steady economic growth and the lowest unemployment rate in decades" (p. 100).

When talking about the electoral success of PiS in connection to social spending pledges made by the party, it should be mentioned that in recent years, PO and PiS have appealed to an increasingly split electorate whose living conditions differ from each other (Przybylski, 2018). According to Przybylski (2018), "While PO developed its constituencies around major urban areas, mostly in the northwest regions of the country, PiS turned to small and medium-sized towns, generally in the southeast" (p. 55). This split encompasses different socioeconomic realities, as economically lower performing regions frequently vote for PiS (Przybylski, 2018). The social spendings made by the PiS government therefore directly affect the living conditions of its core electorate and thus understandably improve the party's standing.

As the previous analysis of the narrative put forward by the PiS government shows, the party places a lot of emphasis on its social spendings and chooses to convey a message emphasising these spendings over one portraying a more ideological narrative. We can connect these actions to efforts to deradicalize the image of the party and appeal to a broader electorate. Furthermore, we should note that there is evidence of political parties becoming less ideologically driven and moving towards a more 'market-oriented' strategy in recent years (Reeves et al., 2006). For example. according to Reeves et al. (2006), political parties in Great Britain increasingly attempt a so-called consumer-driven or market-oriented strategy instead of focusing on party ideology. The concept of political marketing is not a new one (Strömbäck, 2007). According to Strömbäck (2007), the defining features of the concept are:

> *"[First,] [...] that political marketing is the application of marketing principles and procedures–not just marketing techniques and activities–and second, that the processes should be "in response to the needs and wants" of people or groups targeted by the producers–the organizations or campaigns."* (p. 56).

The focus of the PiS government on social spending pledges is a prime example for successful political marketing, as the 500+ programme, as well as the lowering of the retirement age, can be understood as a direct response to the needs of the Polish electorate. By following through on campaign promises that are widely popular and financially feasible (for now), PiS positions itself as an honest advocate for the Polish lower and middle class.

#### Party Leadership.

Jarosław Kaczyński, the head of PiS, has a long track record in Polish political history. In 2001, he founded the party together with his brother Lech Kaczyński, and served as Prime Minister in the PiS-led coalition government from 2005 until 2007 (Grzymala-Busse, 2018). Currently, however, he does not hold any official government positions. In fact, in 2015 it was Kaczyński who selected Andrzej

Duda as the party's presidential candidate and decided that Beata Szydło would serve as Prime Minister of the PiS government (Grzymala-Busse, 2018). Emphasizing a message of "compassionate conservatism" (Fomina & Kucharczyk, 2016, p. 61) under the slogan of 'Good Change', Jarosław Kaczyński was too controversial a personality, too much associated with the authoritarian leanings of PiS, to occupy an official government position (Fomina & Kucharczyk, 2016). This leads to a somewhat odd power structure, with Kaczyński being acknowledged, as Anna Grzymala-Busse phrases it, as the "power behind the throne" (2018, p. 96). To Przybylski (2018), "hiding its most extreme politicians from the public eye" (p. 56) is part of the deradicalization PiS tries to convey to its electorate.

The efforts of PiS to deradicalize its message in order to become more broadly acceptable have already been touched upon earlier. In fact, the strategy of appealing to different subgroups is often pursued as well as rewarded by the electorate (Somer-Topcu, 2015). According to Somer-Topcu (2015), there are several ways through which a political party can broaden its outreach to different groups in society. One of them is to have a variety of positions which represent policy issues from both the left and the right ends of the political spectrum (Somer-Topcu, 2015). Another possible strategy is to not clarify one's political position but rather do the opposite by engaging in a strategy of 'beclouding' (Somer-Topcu, 2015). However, a political party that tries to reach out to a broader electorate does not necessarily need to change its positions: it might be sufficient to change the phrasing of these positions: "[while] the term party strategy is used to refer to actual party behavior, it is not the actual party positions, but voters' perceptions of these positions that influence their vote choice and determine the electoral consequences of a party strategy" (Somer-Topcu, 2015, p. 845). It is hence possible to pursue the strategy of appealing to a broad electorate without changing one's political position. Somer-Topcu (2015) points towards the selection of a specific candidate as another strategy to appear broadly acceptable. In the case of Poland, PiS has managed to do so through the selection of Beata Szydło as their candidate for Prime Minister, who is perceived as more moderate and not as controversial than Jarosław Kaczyński (Fomina & Kucharczyk, 2016; Przybylski, 2018). Indeed, there is evidence suggesting that voters make their decisions based on leading party candidates instead of a party's policies. According to Johnston et al. (2018), research conducted in Great Britain shows that the electoral success of a political party can at least partially be attributed to the electorate's feelings about leading candidates:

> *"Many voters [...] avoid at least some of the effort involved in assembling and assessing information about parties' policies and instead use heuristics such as their feelings about the party leaders as major determinants of their decisions. When party leaders are changed, therefore, differences in voters' feelings about predecessor and successor could lead to changes in party choice."* (p. 2).

The strategic benefit of selecting Beata Szydło as a candidate for the office of Prime Minister is apparent. Her persona, unlike that of Jarosław Kaczyński, is not as firmly linked to the more controversial leanings of PiS (Fomina & Kucharczyk, 2016; Przybylski, 2018). Through her, PiS manages to appeal to a more moderate part of the electorate who would hesitate to vote for Kaczyński (Fomina & Kucharczyk, 2016; Przybylski, 2018). We can, therefore, understand the decision to select Beata Szydło and not Jarosław Kaczyński for the position of the prime minister as another part of a larger strategy to appeal to the general public.

### Migration Policy.

In order to comprehend the importance of PiS's stance on issues of migration, it is important to note that the 2015 election campaign which resulted in PiS's victory was dominated by the European refugee crisis, which was then at its height. (Narkowicz, 2018). Prior to the elections, PO, then in power, had accepted the EU proposition of quotas for a relocation scheme and pledged to accommodate 7000 refugees (Narkowicz, 2018). On the campaign trail, PiS took a particularly tough stance on the refugee question and strongly opposed the idea of the EU relocation scheme, pointing towards Poland's national security and terrorism threats (Narkowicz, 2018). The electorate rewarded the tough stance on the refugee question at the voting booth (Narkowicz, 2018).

Shortly after the 2016 Brussels terrorist attacks, the new Prime Minister Beata Szydło announced that Poland would retract from the EU relocation scheme and not accommodate any migrants (Narkowicz, 2018). The Polish discourse concerning refugees was dominated by the "imagined threat of a Muslim terrorist arriving from Syria disguised as a refugee" (Narkowicz, 2018, p. 358), as the Polish public started to fear terrorist attacks in Poland could be commited by migrants. In addition to the linkage of the refugee question with issues of national security, the discourse was increasingly racialized, with PiS party leader Jarosław Kaczyński suggesting refugees could bring parasites to Poland, a statement that was criticized as being "reminiscent of Nazi rhetoric" (Narkowicz, 2018, p. 366). As shown in the previous chapter, the refugee crisis is used by the PiS government to criticise both the EU and the previous government (Fomina & Kucharczyk, 2016).

The hostile stance of PiS towards accepting refugees is supported by large parts of the Polish electorate. According to a May 2017 CBOS (Centrum Badania Opinii Społecznej; Centre for Public Opinion Research) survey, only 25 percent of Poles were in favour of accepting refugees, while 70 percent were against it (Szczerbiak, 2017). In fact, in a country that as of today is "one of the most religiously and ethnically homogenous countries in Europe with around 90% of the population declaring themselves as Catholics", the 2015 refugees crisis has heightened Islamophobic sentiments and fears concerning national identity (Narkowicz, 2018, p. 359). Issues related to national identity, Catholicism, and conservative values have been emphasized by PiS for years and are at the core of the party's identity (Fomina & Kucharczyk, 2016). In fact, Fomina and Kucharczyk (2016) attribute PiS's "dominant position on the right hand side of the political spectrum and its resilience despite years in opposition" (p. 66) to its strong positions in those areas as well as its connection to the Catholic Church. PiS's tough stance on the question of accepting refugees was therefore credible to the Polish electorate, and especially appealing in a time when concerns about accepting refugees were heightened and considered more pressing than usual. Fomina & Kucharczyk (2016) argue that PiS's "hard line on refugees, verging on xenophobia, won over people who normally would never have voted for PiS" (p. 62). Research shows that a political party's position on a singular issue can indeed influence voters' decisions. As De Sio et al. (2018) elaborate: "A number of studies have shown the increasing importance of the political issues of the day for voting behavior, on both sides of the Atlantic" (p. 1209). De Vries and Hobolt (2012) demonstrate how political parties can actively make use of the importance of singular issues and attract new voters by adopting an 'issue entrepreneurial strategy' (therefore, introducing a new issue dimension). In 2015, PiS was able to put forward a compelling policy standpoint on the refugee crisis: that refugees would not be accepted in Poland. While other parties also adopted positions that reflected the public's anxiety regarding the refugee question, the stance of PiS was the most credible due to the party's history (Narkowicz, 2018). It is therefore indeed possible that a lot of voters who would not have traditionally considered voting for PiS changed their minds because of the refugee crisis (Fomina & Kucharczyk, 2016).

In its dispute with the European Commission over the question of whether to accept refugees, the PiS government managed to position itself as a force protecting Poland from foreign interference (Narkowicz, 2018). While Poland's citizenry is overwhelmingly pro-European (Fomina & Kucharczyk, 2016), many citizens agreed with the PiS government instead of the European Commission when it came to a EU relocation scheme for refugees (Narkowicz, 2018). PiS politicians openly accuse the EU of trying to impose on Poland, with a representative of PiS in the European Parliament, Legutko, "accus[ing] the EU of promoting in this area a "left-liberal" agenda (feminism, LGBT rights, gay marriage, multiculturalism) that must be resisted" (Rupnik, 2018, p. 33).

Furthermore, the dispute with the EU over refugee policy enabled the PiS government to position itself as a victim of unfounded criticism by the EU (Szczerbiak, 2017). This transcends onto other issue areas. When faced with criticism by the Venice Commission concerning the state of the rule of law in Poland, the PiS government invoked a narrative of sovereign democracy in which the EU is yet again presented as a foreign force trying to impose on Polish sovereignty (Przybylski, 2018; Rupnik, 2018). Przybylski (2018) elaborates: "According to this argument, a party that has won the majority of seats in Parliament represents the sovereign will of Poland. Once the election results are in, the new government's democratic legitimacy places its actions above question" (p. 59). The refugee crisis has therefore not only very likely benefitted PiS during the 2015 elections, it also has handed the party a template for dealing with criticism of its actions by the European Union.

## Analysis

### Concrete actions of the PiS government

In order to determine what makes a political party appealing to the electorate, it is important to analyse not only what a party says, but also what a party does (Strömbäck, 2007). The actions of PiS have contributed in various ways to the party's appeal and its domestic support.

First and foremost, PiS's popularity stems from the fact that it delivered on various social spending pledges made in the 2015 campaign. As mentioned before, these promises included a generous child benefit programme as well as the lowering of the retirement age after it had been raised by the previous administration (Fomina & Kucharczyk, 2016). As Poland is currently experiencing a period of economic growth, the PiS government can afford these costly projects, and criticism pointing towards the state of the public finances is largely ignored (Grzymala-Busse, 2018; Szczerbiak, 2017). PiS's focus on social spending can be understood as a form of political marketing, a strategy according to which, simply put, the needs of the market (the electorate) determine the final product (policies) (Strömbäck, 2007). Ideology is no longer the leading factor which determines a voter's decision at the ballot box (Reeves et al., 2006); instead PiS' voters might be persuaded by the successful socioeconomic policies the party has initiated (Fomina & Kucharczyk, 2016).

In addition, voters who in previous years might have been put off by the authoritarian leanings of PiS did not have to fear that the leading candidate of PiS would be too radical. Instead of party leader Jarosław Kaczyński, Beata Szydło was chosen as Prime Minister, effectively broadening the party's appeal by deradicalizing its message (Somer-Topcu, 2015). While Jarosław Kaczyński has long been associated with PiS and continues to lead the party, Szydło was perceived as a less controversial, more moderate candidate (Fomina & Kucharczyk, 2016; Przybylski, 2018). Especially considering the fact that many voters do not necessarily research a political party's platform but are rather guided by their feelings about leading party figures as a heuristic (Johnston et al., 2018), selecting Szydło as Prime Minister was a decision which contributed to broadening PiS's appeal.

Finally, one must point out the importance of the 2015 refugee crisis in order to understand the 2015 electoral victory of PiS. Despite the fact that the country was barely affected by the crisis, the 2015 election was dominated by the topic (Narkowicz, 2018). On this issue, many Poles agreed with the position of PiS which rejected the EU relocation scheme and the notion of accepting Muslim refugees into the country (Szczerbiak, 2017). Research shows that the 'political issue of the day' might indeed be decisive when it comes to a voter's decision (De Sio et al., 2018). It is especially effective when parties manage to explore new issue dimensions and thus explore an issue entrepreneurial strategy in order to win votes (De Vries and Hobolt, 2012). After the new PiS government retracted from the EU relocation scheme, PiS managed to frame itself as the victim of anti-Polish sentiments in an EU that tried to impose a liberal model of society upon Poland (Przybylski, 2018; Rupnik, 2018).

# VIII. Conclusion

In order to understand the strategies PiS uses to appeal to the public, it is necessary to analyze both the narrative the party tries to convey, as well as its concrete actions. A thorough analysis of a 2018 joint press conference by Prime Minister Szydło and PiS President Kaczyński reflects the story the PiS government tells about its own role in Poland. The (until then) two years of PiS governance are represented as a success story, with a special emphasis on the fact that PiS has managed to improve the living conditions of the so-called 'ordinary people'. The conference further highlights the importance of the development of the Polish economy and socioeconomic policies which were particularly popular with the electorate, such as the 500+ programme and the lowering of the pension age to 60 and 65 for women and men, respectively. Government initiatives that were regarded as controversial and/or triggered mass protests in Poland (judicial reform and restrictions on abortion rights) were only briefly mentioned or entirely ignored in the press conference. Another topic which is barely mentioned in the press conference is migration. This is surprising given that large parts of the Polish citizenry agree with PiS on its stances regarding refugees and migrants. The omission of the topic might be due to the fact that questions regarding migration are too closely related to PiS's more radical ideology while the press conference in general attempted to bypass topics traditionally associated with PiS. Another noteworthy aspect is that a narrative of 'reclaiming' and/or 'restoring' the Polish country is repeatedly pushed, insinuating a counterfactual 'Poland in ruins' narrative according to which previous administrations left Poland in shambles (Fomina & Kucharczyk, 2016). While deradicalization of the party's message seems to be a key concern of the official press conference, PiS leaves no doubt that its goal is to change Poland drastically.

In addition to analyzing the official narrative that PiS tries to convey to the Polish citizenry, three significant actions of the PiS government were contextualized and further examined. The party engages in a classic example of political marketing, which becomes especially apparent when analyzing its social spending pledges. Programmes like the 500+ child benefit programme are highly popular with the Polish people and add to the party's appeal. Not only does PiS employ a strategy of softening its image in order to appeal to a broader audience through its narrative, but it also does so through its actions. An additional softening of PiS's image was achieved by selecting Beata Szydło for the position of Prime Minister, who is perceived as less controversial than Jarosław Kaczyński, the PiS party leader (Fomina & Kucharczyk, 2016; Przybylski, 2018). Finally, the rejection of the PiS government to accept refugees added to the party's appeal, as large parts of the Polish citizenry agree with this position and the discourse concerning refugees dominated the 2015 elections. Here, PiS successfully used a 'political issue of the day' to its advantage and subsequently reaped the benefits at the voting booth.

The findings of this research concerning both narrative and concrete actions suggest that PiS has successfully adopted a double strategy in which the party deradicalized its image in order to appeal to new voters while not neglecting its core electorate. The electorate of PiS has historically been located at the far-right of the Polish political spectrum, and its traditional core issues consist of national identity, Catholicism, and conservative values (Fomina & Kucharczyk, 2016). Through its tough stance on migration PiS has kept its campaign pledge to this part of the electorate. One might be tempted to claim that PiS is not engaging in a double strategy, but that its position on the political spectrum has simply shifted to the centre. Is it possible that it was not the message that has been deradicalized, but rather the party? Indeed one must be aware that the character of a political party necessarily changes to at least some extent with the political message it sends. After all, winning over new parts of the electorate subsequently influences PiS's decisions regarding its future political strategies and campaign messages. PiS's success in the *electoral* arena and its changed behaviour in the *media* arena are bound to influence the party's behaviour in the *parliamentary* and the *internal* arena because these arenas overlap (Strömbäck, 2007). However, there is evidence that the radical elements of the party have not changed so much as they have been hidden. PiS's ideology has not changed as much as its political message has; its authoritarian leanings and xenophobic character are still very much existent (Fomina & Kucharczyk, 2016); Kaczyński's position in the party is a prime example of this. In order to appeal to a more moderate public, the key controversial figure does not occupy any governmental positions (Fomina & Kucharczyk, 2016). However, he is still universally acknowledged as the de facto leader of Poland (Fomina & Kucharczyk, 2016; Rupnik, 2018). A substantive change in PiS's ideological makeup would require Kaczyński not to step out of the spotlight, but to step down as the party's leader. An analysis of the party's concrete actions, therefore, results in the conclusion that the party is engaging in a double strategy which is changing its message in order to appeal to a mode moderate public, while still not losing its core electorate, by maintaining a more radical image for those 'in the know'.

A greater variety of sources would enable this research project to have a more holistic analysis. It is presumed that PiS deradicalizes its message depending on the audience it targets. The press conference from which the narrative was derived was a highly official event (both Prime Minister and party leader participated, its occasion was the celebration of two years of successful governmental rule, and it was translated into English for international observants) from which one could expect a softening of PiS's message. It is thinkable that the double strategy in which PiS is currently engaging would not have been recognized if this research had analyzed content solely from a different, e.g. more local, context. To do so is a task for future research in the topic area.

While PiS has managed to convey a softer, less radical message to voters, its ideological core with its authoritarian and xenophobic leanings is still very much existent, as Kaczyński's position as party leader suggests. However, the very fact that PiS has engaged in an act of deradicalizing its message shows that the party is changing. By attracting new subgroups of voters, the identity of the party is not drastically altered, but certainly has become less clear-cut. Today, not everyone who identifies with PiS necessarily identifies with the traditional notion of conservative values the party was built on. For Poland, this situation could pose a chance as well as a threat. If PiS were to abandon its authoritarian aspirations, its double strategy could further a transformation of the PiS into a Christian conservative people's party without extremist tendencies, for example. Such a party could possibly fill the space in the Polish political spectrum formerly occupied by PO. However, the current behaviour of PiS, especially in the realm of judicial reform, makes this scenario improbable. It seems more likely that PiS will continue to systematically undermine institutions of liberal democracy, entailing the courts and the media. Poland's transformation into an illiberal democracy will continue until a substantive part of the Polish citizenship recognizes this transition as a threat.

On the European level, this means that illiberal movements across the continent will be strengthened. This is especially true in, but not limited to, the CEE region and Orbán's Fidesz party in Hungary. Future research on the political strategies of parties comparable to PiS should determine to what extent the success of PiS can be attributed to structures that are specific to the Polish context. Political movements similar to PiS are found across national and even continental borders (Rupnik, 2018), and in order to fully comprehend their root causes, it is of utmost importance to understand what makes them more successful in some countries than in others.

## Works Cited

Brysk, A. (1995). " Hearts and minds": bringing symbolic politics back in. *Polity*, 27(4), 559-585.

Chancellery of the Prime Minister (2017). *Prime Minister Beata Szydło: family, development and security are the three pillars of our governance* [Press release]. Retrieved April 21, 2019, from https://www.premier.gov.pl/mobile/en/news/news/prime-minister-beata-szydlo-family-development-and-security-are-the-three-pillars-of-our.html

Cianetti, L., Dawson, J., & Hanley, S. (2018). Rethinking "democratic backsliding" in Central and Eastern Europe – looking beyond Hungary and Poland. *East European Politics*, 34(3), 243–256. https://doi.org/10.1080/21599165.2018.1491401

De Sio, L., De Angelis, A., & Emanuele, V. (2018). Issue yield and party strategy in multiparty competition. *Comparative Political Studies*, 51(9), 1208-1238.

De Vries, C. E., & Hobolt, S. B. (2012). When dimensions collide: The electoral success of issue entrepreneurs. *European Union Politics*, 13(2), 246-268.

Fomina, J., & Kucharczyk, J. (2016). Populism and Protest in Poland. *Journal of Democracy*, 27(4), 58–68. https://doi.org/10.1353/jod.2016.0062

Grzymala-Busse, A. (2018). Poland's Path to Illiberlism. *Current History*, 117(797), 96-101.

Jaskiernia, J. (2017). The Development of the Polish Party System. *Polish Political Science Yearbook*, 46(2), 227-246.

Johnston, R., Hartman, T., & Pattie, C. (2018). Feelings about party leaders as a voter's heuristic–What happens when the leaders change? A note. *Electoral Studies*.

Kowalewski, O., & Rybinski, K. (2011). The hidden transformation: the changing role of the state after the collapse of communism in Central and Eastern Europe. *Oxford Review of Economic Policy*, 27(4), 634-657.

Przybylski, W. (2018). Can Poland's Backsliding Be Stopped? *Journal of Democracy*, 29(3), 52–64. https://doi.org/10.1353/jod.2018.0044

Mucha, J. (2006). Poland in Central and Eastern Europe, Polish Sociology within the Central European Context: Introduction. *Journal of Classical Sociology*, 6(3), 251-256.

Narkowicz, K. (2018). ‘Refugees Not Welcome Here’: State, Church and Civil Society Responses to the Refugee Crisis in Poland. *International Journal of Politics, Culture, and Society*, 31(4), 357-373.

O’Neil, P. H. (2015). *Essentials of comparative politics: Fifth international student edition*. WW Norton & Company.

Patterson, M., & Monroe, K. R. (1998). Narrative in political science. *Annual review of political science*, 1(1), 315-331. Przybylski, W. (2018). Can Poland’s Backsliding Be Stopped?. *Journal of Democracy*, 29(3), 52-64.

Reeves, P., de Chernatony, L., & Carrigan, M. (2006). Building a political brand: Ideology or voter-driven strategy. *Journal of Brand Management*, 13(6), 418-428.

Rupnik, J. (2018). The Crisis of Liberalism. *Journal of Democracy*, 29(3), 24–38. https://doi.org/10.1353/jod.2018.0042

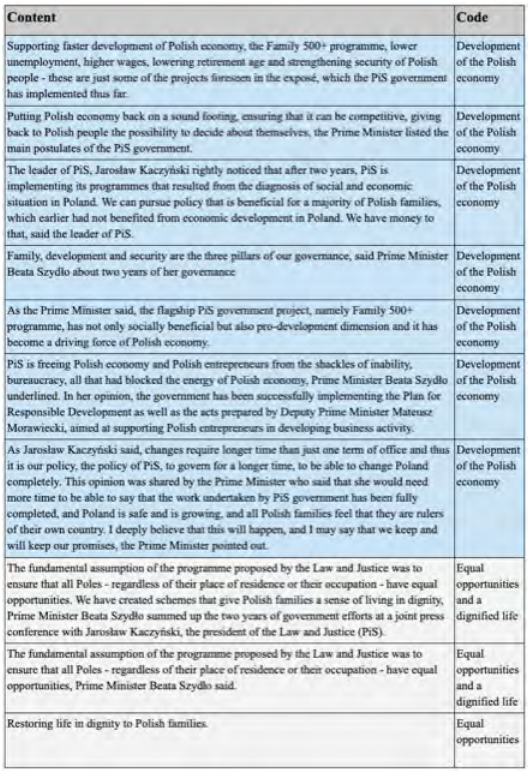
Somer-Topcu, Z. (2015). Everything to everyone: The electoral consequences of the broad-appeal strategy in Europe. *American Journal of Political Science*, 59(4), 841-854.

Strömbäck, J. (2007). Political marketing and professionalized campaigning: A conceptual analysis. *Journal of political marketing*, 6(2-3), 49-67.

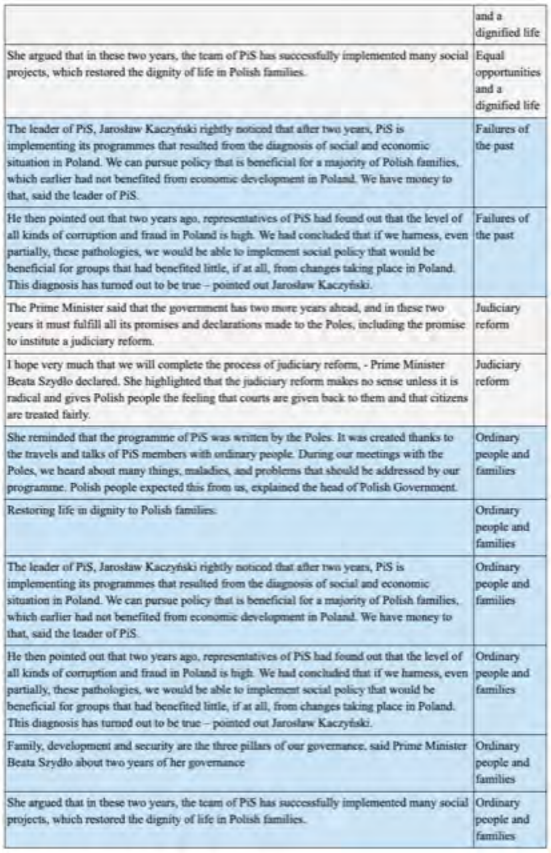
Szczerbiak, A. (2017, November 01). Explaining the popularity of Poland’s Law and Justice government [web log comment]. Retrieved May 5, 2019, from https://blogs.lse.ac.uk/europpblog/2017/10/26/explaining-the-popularity-of-polands-law-and-justice-government/

Waldner, D., & Lust, E. (2018). Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*, 21, 93-113.

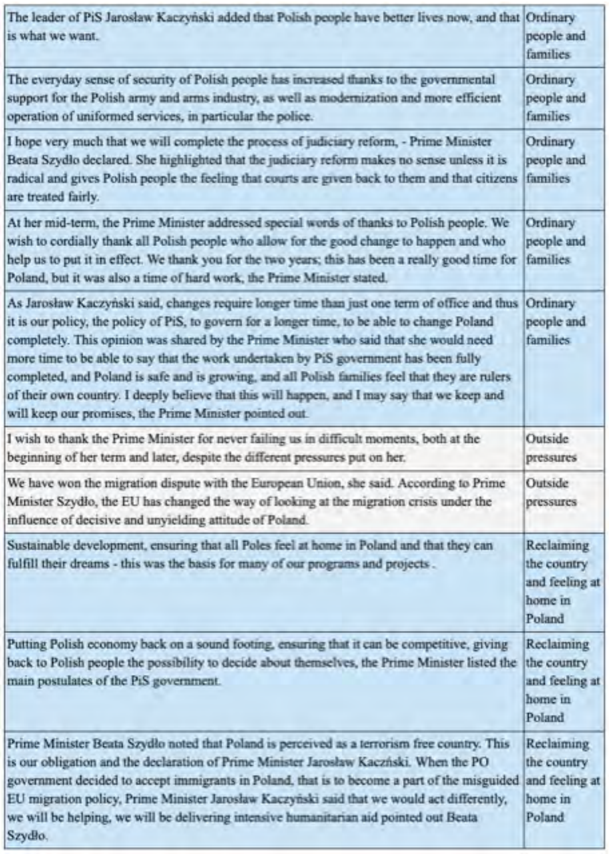
# Appendix A



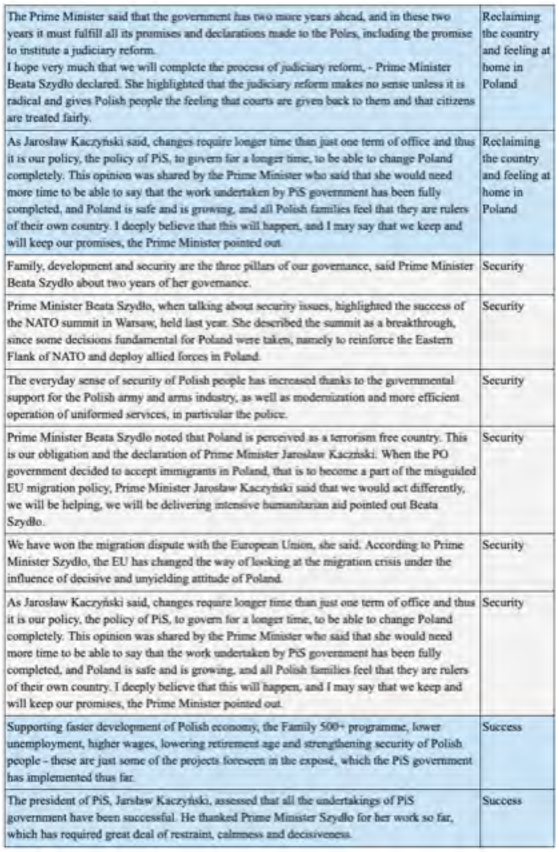
| Content | Code |
|---|---|
| Supporting faster development of Polish economy, the Family 500+ programme, lower unemployment, higher wages, lowering retirement age and strengthening security of Polish people - these are just some of the projects foreseen in the exposé, which the PiS government has implemented thus far. | Development of the Polish economy |
| Putting Polish economy back on a sound footing, ensuring that it can be competitive, giving back to Polish people the possibility to decide about themselves, the Prime Minister listed the main postulates of the PiS government. | Development of the Polish economy |
| The leader of PiS, Jarosław Kaczyński rightly noticed that after two years, PiS is implementing its programmes that resulted from the diagnosis of social and economic situation in Poland. We can pursue policy that is beneficial for a majority of Polish families, which earlier had not benefited from economic development in Poland. We have money to that, said the leader of PiS. | Development of the Polish economy |
| Family, development and security are the three pillars of our governance, said Prime Minister Beata Szydło about two years of her governance | Development of the Polish economy |
| As the Prime Minister said, the flagship PiS government project, namely Family 500+ programme, has not only socially beneficial but also pro-development dimension and it has become a driving force of Polish economy. | Development of the Polish economy |
| PiS is freeing Polish economy and Polish entrepreneurs from the shackles of inability, bureaucracy, all that had blocked the energy of Polish economy, Prime Minister Beata Szydło underlined. In her opinion, the government has been successfully implementing the Plan for Responsible Development as well as the acts prepared by Deputy Prime Minister Mateusz Morawiecki, aimed at supporting Polish entrepreneurs in developing business activity. | Development of the Polish economy |
| As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely. This opinion was shared by the Prime Minister who said that she would need more time to be able to say that the work undertaken by PiS government has been fully completed, and Poland is safe and is growing, and all Polish families feel that they are rulers of their own country. I deeply believe that this will happen, and I may say that we keep and will keep our promises, the Prime Minister pointed out. | Development of the Polish economy |
| The fundamental assumption of the programme proposed by the Law and Justice was to ensure that all Poles - regardless of their place of residence or their occupation - have equal opportunities. We have created schemes that give Polish families a sense of living in dignity, Prime Minister Beata Szydło summed up the two years of government efforts at a joint press conference with Jarosław Kaczyński, the president of the Law and Justice (PiS). | Equal opportunities and a dignified life |
| The fundamental assumption of the programme proposed by the Law and Justice was to ensure that all Poles - regardless of their place of residence or their occupation - have equal opportunities, Prime Minister Beata Szydło said. | Equal opportunities and a dignified life |
| Restoring life in dignity to Polish families. | Equal opportunities |



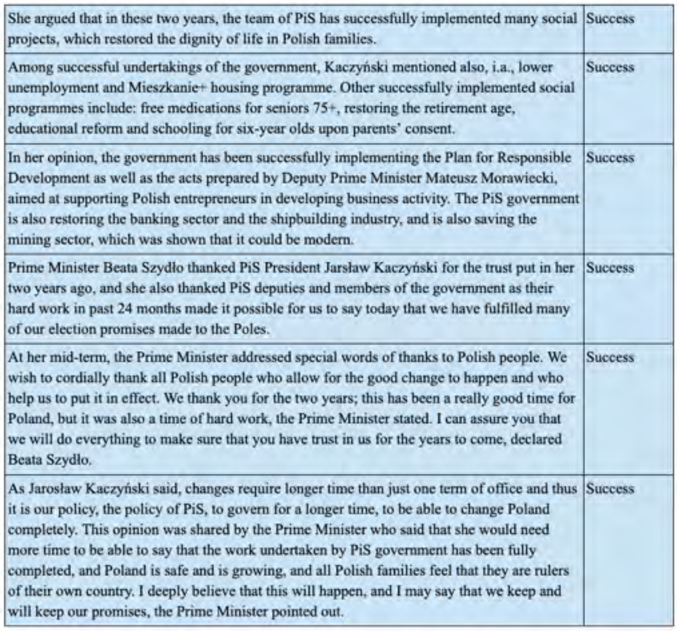
| | |
|---|---|
| | and a dignified life |
| She argued that in these two years, the team of PiS has successfully implemented many social projects, which restored the dignity of life in Polish families. | Equal opportunities and a dignified life |
| The leader of PiS, Jarosław Kaczyński rightly noticed that after two years, PiS is implementing its programmes that resulted from the diagnosis of social and economic situation in Poland. We can pursue policy that is beneficial for a majority of Polish families, which earlier had not benefited from economic development in Poland. We have money to that, said the leader of PiS. | Failures of the past |
| He then pointed out that two years ago, representatives of PiS had found out that the level of all kinds of corruption and fraud in Poland is high. We had concluded that if we harness, even partially, these pathologies, we would be able to implement social policy that would be beneficial for groups that had benefited little, if at all, from changes taking place in Poland. This diagnosis has turned out to be true – pointed out Jarosław Kaczyński. | Failures of the past |
| The Prime Minister said that the government has two more years ahead, and in these two years it must fulfill all its promises and declarations made to the Poles, including the promise to institute a judiciary reform. | Judiciary reform |
| I hope very much that we will complete the process of judiciary reform, - Prime Minister Beata Szydlo declared. She highlighted that the judiciary reform makes no sense unless it is radical and gives Polish people the feeling that courts are given back to them and that citizens are treated fairly. | Judiciary reform |
| She reminded that the programme of PiS was written by the Poles. It was created thanks to the travels and talks of PiS members with ordinary people. During our meetings with the Poles, we heard about many things, maladies, and problems that should be addressed by our programme. Polish people expected this from us, explained the head of Polish Government. | Ordinary people and families |
| Restoring life in dignity to Polish families. | Ordinary people and families |
| The leader of PiS, Jarosław Kaczyński rightly noticed that after two years, PiS is implementing its programmes that resulted from the diagnosis of social and economic situation in Poland. We can pursue policy that is beneficial for a majority of Polish families, which earlier had not benefited from economic development in Poland. We have money to that, said the leader of PiS. | Ordinary people and families |
| He then pointed out that two years ago, representatives of PiS had found out that the level of all kinds of corruption and fraud in Poland is high. We had concluded that if we harness, even partially, these pathologies, we would be able to implement social policy that would be beneficial for groups that had benefited little, if at all, from changes taking place in Poland. This diagnosis has turned out to be true – pointed out Jarosław Kaczyński. | Ordinary people and families |
| Family, development and security are the three pillars of our governance, said Prime Minister Beata Szydło about two years of her governance | Ordinary people and families |
| She argued that in these two years, the team of PiS has successfully implemented many social projects, which restored the dignity of life in Polish families. | Ordinary people and families |



| | |
|---|---|
| The leader of PiS Jarosław Kaczyński added that Polish people have better lives now, and that is what we want. | Ordinary people and families |
| The everyday sense of security of Polish people has increased thanks to the governmental support for the Polish army and arms industry, as well as modernization and more efficient operation of uniformed services, in particular the police. | Ordinary people and families |
| I hope very much that we will complete the process of judiciary reform, - Prime Minister Beata Szydło declared. She highlighted that the judiciary reform makes no sense unless it is radical and gives Polish people the feeling that courts are given back to them and that citizens are treated fairly. | Ordinary people and families |
| At her mid-term, the Prime Minister addressed special words of thanks to Polish people. We wish to cordially thank all Polish people who allow for the good change to happen and who help us to put it in effect. We thank you for the two years; this has been a really good time for Poland, but it was also a time of hard work, the Prime Minister stated. | Ordinary people and families |
| As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely. This opinion was shared by the Prime Minister who said that she would need more time to be able to say that the work undertaken by PiS government has been fully completed, and Poland is safe and is growing, and all Polish families feel that they are rulers of their own country. I deeply believe that this will happen, and I may say that we keep and will keep our promises, the Prime Minister pointed out. | Ordinary people and families |
| I wish to thank the Prime Minister for never failing us in difficult moments, both at the beginning of her term and later, despite the different pressures put on her. | Outside pressures |
| We have won the migration dispute with the European Union, she said. According to Prime Minister Szydło, the EU has changed the way of looking at the migration crisis under the influence of decisive and unyielding attitude of Poland. | Outside pressures |
| Sustainable development, ensuring that all Poles feel at home in Poland and that they can fulfill their dreams - this was the basis for many of our programs and projects . | Reclaiming the country and feeling at home in Poland |
| Putting Polish economy back on a sound footing, ensuring that it can be competitive, giving back to Polish people the possibility to decide about themselves, the Prime Minister listed the main postulates of the PiS government. | Reclaiming the country and feeling at home in Poland |
| Prime Minister Beata Szydło noted that Poland is perceived as a terrorism free country. This is our obligation and the declaration of Prime Minister Jarosław Kaczński. When the PO government decided to accept immigrants in Poland, that is to become a part of the misguided EU migration policy, Prime Minister Jarosław Kaczyński said that we would act differently, we will be helping, we will be delivering intensive humanitarian aid pointed out Beata Szydło. | Reclaiming the country and feeling at home in Poland |



| | |
|---|---|
| The Prime Minister said that the government has two more years ahead, and in these two years it must fulfill all its promises and declarations made to the Poles, including the promise to institute a judiciary reform.<br>I hope very much that we will complete the process of judiciary reform, - Prime Minister Beata Szydlo declared. She highlighted that the judiciary reform makes no sense unless it is radical and gives Polish people the feeling that courts are given back to them and that citizens are treated fairly. | Reclaiming the country and feeling at home in Poland |
| As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely. This opinion was shared by the Prime Minister who said that she would need more time to be able to say that the work undertaken by PiS government has been fully completed, and Poland is safe and is growing, and all Polish families feel that they are rulers of their own country. I deeply believe that this will happen, and I may say that we keep and will keep our promises, the Prime Minister pointed out. | Reclaiming the country and feeling at home in Poland |
| Family, development and security are the three pillars of our governance, said Prime Minister Beata Szydlo about two years of her governance. | Security |
| Prime Minister Beata Szydło, when talking about security issues, highlighted the success of the NATO summit in Warsaw, held last year. She described the summit as a breakthrough, since some decisions fundamental for Poland were taken, namely to reinforce the Eastern Flank of NATO and deploy allied forces in Poland. | Security |
| The everyday sense of security of Polish people has increased thanks to the governmental support for the Polish army and arms industry, as well as modernization and more efficient operation of uniformed services, in particular the police. | Security |
| Prime Minister Beata Szydlo noted that Poland is perceived as a terrorism free country. This is our obligation and the declaration of Prime Minister Jaroslaw Kaczniski. When the PO government decided to accept immigrants in Poland, that is to become a part of the misguided EU migration policy, Prime Minister Jarosław Kaczyński said that we would act differently, we will be helping, we will be delivering intensive humanitarian aid pointed out Beata Szydło. | Security |
| We have won the migration dispute with the European Union, she said. According to Prime Minister Szydlo, the EU has changed the way of looking at the migration crisis under the influence of decisive and unyielding attitude of Poland. | Security |
| As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely. This opinion was shared by the Prime Minister who said that she would need more time to be able to say that the work undertaken by PiS government has been fully completed, and Poland is safe and is growing, and all Polish families feel that they are rulers of their own country. I deeply believe that this will happen, and I may say that we keep and will keep our promises, the Prime Minister pointed out. | Security |
| Supporting faster development of Polish economy, the Family 500+ programme, lower unemployment, higher wages, lowering retirement age and strengthening security of Polish people - these are just some of the projects foreseen in the exposé, which the PiS government has implemented thus far. | Success |
| The president of PiS, Jarsław Kaczyński, assessed that all the undertakings of PiS government have been successful. He thanked Prime Minister Szydlo for her work so far, which has required great deal of restraint, calmness and decisiveness. | Success |



| | |
|---|---|
| She argued that in these two years, the team of PiS has successfully implemented many social projects, which restored the dignity of life in Polish families. | Success |
| Among successful undertakings of the government, Kaczyński mentioned also, i.a., lower unemployment and Mieszkanie+ housing programme. Other successfully implemented social programmes include: free medications for seniors 75+, restoring the retirement age, educational reform and schooling for six-year olds upon parents' consent. | Success |
| In her opinion, the government has been successfully implementing the Plan for Responsible Development as well as the acts prepared by Deputy Prime Minister Mateusz Morawiecki, aimed at supporting Polish entrepreneurs in developing business activity. The PiS government is also restoring the banking sector and the shipbuilding industry, and is also saving the mining sector, which was shown that it could be modern. | Success |
| Prime Minister Beata Szydło thanked PiS President Jarsław Kaczyński for the trust put in her two years ago, and she also thanked PiS deputies and members of the government as their hard work in past 24 months made it possible for us to say today that we have fulfilled many of our election promises made to the Poles. | Success |
| At her mid-term, the Prime Minister addressed special words of thanks to Polish people. We wish to cordially thank all Polish people who allow for the good change to happen and who help us to put it in effect. We thank you for the two years; this has been a really good time for Poland, but it was also a time of hard work, the Prime Minister stated. I can assure you that we will do everything to make sure that you have trust in us for the years to come, declared Beata Szydło. | Success |
| As Jarosław Kaczyński said, changes require longer time than just one term of office and thus it is our policy, the policy of PiS, to govern for a longer time, to be able to change Poland completely. This opinion was shared by the Prime Minister who said that she would need more time to be able to say that the work undertaken by PiS government has been fully completed, and Poland is safe and is growing, and all Polish families feel that they are rulers of their own country. I deeply believe that this will happen, and I may say that we keep and will keep our promises, the Prime Minister pointed out. | Success |

Humanities - Social Sciences

# *Between the World and Me(me):* Approaching Internet Memes through Digital Ethnography and Berlant's Affective Present

Cullen Ogden

*Supervisor*
Dr. Alexandra Brown (AUC, UvA)
*Reader*
Dr. Pedram Dibazar (AUC)

**Abstract**

This research addresses the massive number of internet memes encountered daily and over time to consider how the technological capabilities of the internet have been able to open new avenues for relationality and subjectivity, fundamentally shaping how we perceive what's true, what's right, and what's possible. By combining a two-step methodology of gathering auto-ethnographic, qualitative research and then analyzing this data through Lauren Berlant's theory of the contemporary affective present, this project argues that memes constitute a genre of the contemporary present, a subgenre of comedy made possible by digital technology. From this claim of the meme as a genre, this research extends Berlant's theory by developing the terms *self-connection* and *virtual intuition* to speak to the respective (re)shaping of subjectivity and intuition today.

Keywords and phrases: *everyday life, memes, Lauren Berlant, affective present, genre*

## Acknowledgements

Thank you to Dr. Robert Y. Chang, my first anthropology professor who taught me the magic of questioning the world around us, and who showed me that a study of culture can encompass any interest you can think of. In addition, thank you to Dr. Pedram Dibazar, who helped me discover the beauty of work done between anthropology and cultural analysis, who has been helpful and supportive not only as a professor but as a friend. I would also like to extend an enormous thank you to Dr. Alexandra Brown, my supervisor, whose support through this entire process has been so formative that I would have never considered a project of this nature without her guidance, and to whom I'm thankful for introducing me to the academic work that continues to grab my attention.

In addition, to my parents, Carolyne and Chandler Ogden, thank you for constantly supporting me in everything I do and encouraging me in all of my oddity. To Manar and the Cult of friends (who feel like family) that I've made in life who have fearlessly supported me, who have been there for me at every step of this process, and who have heard me attempt to explain that wild and wonderful thing that is "affect" too many times to count – Thank you.

And last, but certainly not least, thank you to Lauren Berlant, whose work has been a guiding force in this project, and who taught me the power and potential of articulating that *something* to which we all strive for.

## Introduction:

### *Encountering Berlant*

Lauren Berlant's *Cruel Optimism* is a groundbreaking demonstration of the ambitious and powerful work that can be done through considering the role of affect in the shaping and apprehension of the world today (2011, p. 4). Berlant asserts that, "the present is what makes itself present to us before it becomes anything else, such as an orchestrated collective event or an epoch on which we can look back" (p. 4). This is to say that although the present has substance which we firstly feel, this present is molded by the forms we give to it — forms both temporal and aesthetic which can carry a profoundly political purpose. By assigning the present such forms as a crisis, an interruption, a situation, or a habit, these forms each shape and alter "what forces should be considered responsible and what crises urgent in our adjudication of survival strategies and conception of a better life than what the metric of survival can supply" (Berlant, 2011, p. 4).

Forms matter because today, life has changed and feels particularly precarious and turbulent. This sense of precarity is not a new condition in the world, but its growing presence in the lives of those previously unaffected has generalized this condition to be perceived as an overarching tone of current conditions at large, whose effects and distribution remains unequal (Berlant, as found in Puar, 2012). To consider what this precarious world of today may feel like, Berlant theorizes the present as punctuated by a sense of crisis ordinariness in which the world is *disorganized* by capitalism and other forces that demand the subject to scramble for new ways of living, “rhythms that could, at any time, congeal into norms, forms, and institutions” (2011, p. 9). Part of what reinforces Berlant’s claim to the disorganization of the world is the fraying of what she sees as the “moral-intimate-economic thing called ‘the good life,’” or the conventional fantasies of what it means to have a life, what this survival is for, and how the world “add[s] up to something” (2011, p. 2). As the system appears not to have radically changed but rather increasingly fails to reciprocate the promises of life before in its unchanging nature, these fantasies are fraying. Such ‘good life’ fantasies include:

> “upward mobility, job security, political and social equality... lively, durable intimacy... meritocracy, the sense that liberal-capitalist society will reliably provide opportunities to carve out relations of reciprocity that seem fair and that foster life as a project of adding up to something and constructing cushions for enjoyment”

(Berlant, 2011, p. 3).

The title of her work, *Cruel Optimism*, thus reflects a general affective structure found throughout life today; we, as people in this precarious world, remain attached to objects which we believe build towards this ‘good life’ even as these objects no longer aide us in achieving these fantasies and in fact inhibit our achievement in the process of attachment (Berlant, 2011, p. 1). Specifically, *Cruel Optimism* is about how ordinary life is home to these conflicting forces of old fantasies and new disappointing realities, and how in renegotiating what it means to have and live a life, we find that adapting to and surviving life today may be as close to this good life as we can hope for (Berlant, 2011, p. 3). While to speak of cruel optimism and more generally of affective structure is not to speak directly to the experience of the structure, but to recognize how the act of giving shape to an amorphous ongoing present has effects and consequences. Description then also becomes *prescriptive*: the difference, for example, between ‘climate *change*’ and ‘climate *crisis*’ is neither neutral nor coincidental, and conditions how we adapt and adjust to these forms of the present. By exploring the many different forms and genres which can conflict, contest, and challenge each other, we can ground this study of affect in material narratives and stories which emerge as genres. Focusing on these newly emerging genres gives us a way to consider how the contemporary present is understood and encountered as it unfolds.

So what constitutes cruel optimism in the present? What everyday actions, genres, and practices enact the affective structure Berlant identifies? This capstone explores the relevance of Berlant’s theory in examining one of the most simultaneously prolific and quotidian elements of the contemporary present, which I will argue provides a new genre for apprehending and experiencing the present today: the meme [1].

> “I feel all sorts of ways – angry at a Trump tweet, laughing at a meme about anxiety but also feeling shame, I find another account which enjoys astrology and I follow them which is exciting, and then I laugh but I’m sad and worried about the environment. My thoughts aren’t everywhere, exactly, but they’re many places which I’m entering and moving between within a single screen. But within this chaos, I’m still learning and laughing and imagining absurd situations and a lot of funny things, and I’m learning new slang and new jokes and new ways of relating in themselves. Maybe I retweet some of the ones I find really funny, I definitely like the ones I find funny even if I don’t retweet them, but many I may just scroll past.”

(observation, March 11th, 2019)

The contemporary circulation of and engagement with memes, just as Berlant’s affective present, involves: (i) the setting of expectations, (ii) the shaping and reshaping of the present, and in particular, (iii) an identifiable affective structure which reflects particular facets of the contemporary present that

[1] This research focuses on the Internet Meme specifically, so the use of the word meme is with the assumption that we are discussing memes as they occur online. See Chapter 1 for a working definition of meme.

Berlant describes. While *Cruel Optimism* takes a variety of cases from between 1990 and 2010 which capture this titular affective structure, this project extends Berlant's affective present to include social media and digital technology in the emergence of new genres. Considering the meme in its context of social media and everyday life requires new language which can capture the dynamism and flux behind the term 'meme.'

To think of memes within the context of the affective present is to allow the meme to speak to and from everyday experience, by identifying patterns across a variety of memes. What is a meme at its simplest? In considering how the present is the space where we encounter and gather various forms of knowledge and develop intuition about what's happening and what might follow from this, what kind of knowledge and intuition do memes require? How does the sheer volume of memes under continual production reflect the contemporary present, as identified by Berlant? An interdisciplinary approach that combines digital ethnography, autoethnography, affect theory, and literary criticism will be used to consider what and how we can learn about today's world through memes.

## Project Methodology

The aim of this project is to understand the meme by considering what an affective approach to the meme can begin to tell us about memes, ourselves, and our world. To accomplish this, Berlant's formalist approach to locating affect within cultural objects and discerning from them a sense of the historical present lets us place memes at the intersection of the everyday and the larger contemporary present. Through the emulation, reproduction, and potential habituation of the meme form(s), genres are shared and negotiated as active agents in shaping our perception of the present, which establishes the gravity of these funny little things. Berlant's theory is productive for an analysis of memes, especially in considering how they function and move within everyday life (which will be further established in Chapter 1). However, her own methodology of generalization through reading cultural artifacts proves difficult to apply to a single meme or to the blurry definition of memes as a whole. Thus this paper employs methods of digital anthropology as an evidential basis for interpreting memes through the theoretical framework offered by Berlant's concept of the affective present. The digital ethnographic work produces an extensive archive of instances from which to deduce genres without requiring a classificatory definition of the meme to haunt over the fieldwork process.

Thus, this interdisciplinary research project firstly collected qualitative, auto-ethnographic data from engaged participant observation on social media networks to survey memes which were circulated between March and May of 2019. Taking the everyday act of scrolling on one's phone and looking at memes as an (auto-)ethnographic fieldsite, my fieldwork consisted of three visits of two hours each via laptop computer and mobile phone to the social platforms of Twitter and Instagram, which took place on March 11th, April 3rd, and May 4th, 2019. Through a trace ethnographic approach to engaging with memes as digital 'documents' within the Internet 'network,' this process involved deep dives online, wherein I would emulate the everyday practices of browsing and scrolling to encounter and engage with memes in a variety of forms and digital locations (Geiger Ribes, 2011). Although it would seem digital worlds would require new ways of understanding what it means to participate in comparison to classical anthropology, Tom Boellstorff's work *Ethnography and Virtual Worlds* provides insight into the validity of participant observation both on- and offline, or wherever sociality may be occurring (Boellstorff, 2012, p. 65). The work of Larissa Hjorth, John Postill, and Sarah Pink has also been impactful in demonstrating innovative approaches to considering digital cultures through ethnographic engagement.

During the ethnographic work, memes were encountered not only via my own networks, but also by utilizing tags and searching features to see what the social media algorithms of Twitter and Instagram might produce – behaviors potentially exhibited by those who have yet to discover what a 'meme' is. Other collection methods included acts such as writing posts that asked explicitly for memes, messaging a number of friends asking for memes they recently came across, and predominantly autonomous navigating of these networks. By using my own social media accounts, my own experience of these encounters was already filtered through the networks of followers and following that I had developed on each platform respectively. To prevent reliance on my self-curated timelines, I utilized the interactive aspects of new media to follow the trails of sociality through the interface instead. Additionally, documenting the entire process through fieldnotes has provided an archive to revisit in further considering the development of this argument.

In the second step of this interdisciplinary project, I incorporated Berlant's theory of genre and the affective present to code and analyze my findings from my digital ethnographic exploration. In addition to concepts developed in *Cruel Optimism*, Berlant's additional work on precarity and comedy provide useful supplements to this framework of genre. Berlant develops a broad notion of genre to describe any affective structure which trains our expectations of watching something unfold, a useful conceptual tool for an affective analysis which will be further examined in Chapter 2 (2011, p. 7). As the first chapter will describe, Berlant's work provides a relevant and provocative framework for thinking about everyday life in the contemporary present by highlighting how genre works to (re)frame and (re)shape the present as and after being perceived.

Ultimately, this research argues that memes constitute a genre of the contemporary present, operating within temporal genres which have been enabled and heightened through digital technologies. As a genre, memes therefore function in multiple ways: first, by developing new modes of subjectivity through engagement with memes via social media; and secondly, by heightening our sense of intuition whereby we come to relate in new ways to the world and its forces.

In Chapter 1, this research reviews how memes and everyday life (via anthropology and everyday life theory) have been studied in the past to contextualize Berlant's and my own work. In Chapter 2, I define and elaborate on Berlant's concept of genre and introduce the genre of comedy to consider how memes extend a particular form of comedy into the digital. From this definition, the last two chapters take up the meme as genre to consider the effects and implications of this structure in shaping this affective present. Chapter 3 considers how the meme affects subjectivity, while Chapter 4 will address how the meme effects intuition. Finally, the conclusion will reflect on this research and the work of theorizing comedy.

## I. Memes Among Us: Memes and/in Everyday Life

Limor Shifman, a predominant contemporary scholar on memes, defines an internet meme as, "(a) a *group of digital items sharing common characteristics* of content, form, and/or stance; (b) that were *created with awareness of each other*; and (c) were circulated, imitated, and/or transformed *via the Internet by many users*" (Shifman, 2014, p. 8). While this definition draws on the use of the term 'internet' to develop a definition of the object in relation to its medium (the Internet network), this does not consider the ambiguity of the term as such and the confusion in defining its multiplicity. To answer what the 'meme' at the heart of the 'internet meme' is requires demystifying the term 'meme' and its double meaning within colloquial common knowledge and academic use.

### Memes

*Meme* was first coined by Richard Dawkins in 1976 as a biological analogy applied to culture, but only started to gain pop-culture traction after Mike Godwin published an article about Nazi memes in *Wired* magazine in 1993 (Shifman, 2014, p. 1; Godwin, 1993). Defined loosely here as 'units of culture,' memes arose as a tool for exploring how the transmission of culture and particular cultural ideas may be similar to that of genes, or at least involve similar processes such as transmission, duplication, and mutation. Some eagerly jumped on board to this pursuit, with even the field of 'memetics' briefly developing in hopes of establishing the study of memes in academia. However, critics of this discipline have been unable to reconcile the flawed origins of the cultural-biological metaphor, with worries over agency and transposing scientific theory onto culture (Shifman, 2014, p. 10; Jenkins et al., 2009; Conte, 2000). Many who have worked on memes and confronted these issues have chosen to simplify the unruly meme by considering it within a specific disciplinary context. This has illuminated only a small piece of the dynamic processes involved in the production of even a single meme, and isolated the meme from its context in circulation.

This can be seen in a variety of fields as memes are conceptualized within a specific discipline, and have been considered in the domains of philosophy (Sterelny, 2006; Holdcraft Lewis, 2000), psychology (Blackmore, 1999; Blackmore, 2001), anthropology (Marwick, 2013; Shifman, 2014; Nissenbaum Shifman, 2015; Banet-Weiser Miltner, 2015; Saito, 2017), linguistics (Zenner Geeraerts, 2018), politics (Borenstein, 2004; Coker, 2008; Dağtaş, 2016), as well as communication studies and aesthetics (Dennett, 1990; Rossolatos, 2015; Gal et al., 2016). My concern with these previous approaches is that investigating a specific meme from a specific context requires assuming a stance toward memes that can distort or generalize particular cases in efforts to fulfill the aim of the discipline at large. A solution to this could be to compile these different investigations to produce a fuller image of what an internet meme is. On the contrary, one could simply claim that all studies of memes are destined to fail due to the impossibility to ever include all memes and this potential for a meme to contradict any findings. Alternatively, I would like to ask: what if we instead framed memes as a part of the "residual of life," that excess which cannot be captured by simply one single approach but instead requires new paradigms for consideration (Highmore, 2008, p. 3)? What if we position ourselves to academically approach memes in the same way we personally encounter them, not in isolation but instead in the chaos within which we happen upon them – within the everyday?

## Everyday Life

To turn to the everyday as a means of interpreting the meme is not an easy task, but cultural studies and anthropology offer two different yet complementary trajectories for theorizing this. Considering first the cultural studies approach to everyday life upon which everyday life theory has been constructed, the everyday develops "a kind of heuristic approach to social life that does not start out with predesignated outcomes" (Highmore, 2008, p. 3). While the study of everyday life became more explicit in the latter half of the 20th century, Highmore claims that the origins of turning to everyday life first appeared in the works of Sigmund Freud and Karl Marx (2014, p. 6). Through a framework of skepticism, Freud and Marx planted the seeds to challenge the regularizing and totalizing power of theory in producing notions of agency, determinism, and rationality that had been present during the 19th century, and in fostering critical perspectives on everyday life (Highmore, 2014, p. 6).

One of the first scholars to directly engage with the conceptual realm of the everyday and attempt to theorize it was Henri Lefebvre. In his *Critique of Everyday Life*, he claimed that everyday life is defined by contradictions and stands at the magical intersection of "illusion and truth, power and helplessness; the intersection of the sector man controls and the sector he does not control" (Lefebvre, 1992, p. 21). His critique of everyday life took issue with the different rhythms that all come together in the everyday, and he argued that by recognizing these conflicting forces we would come to recognize that the everyday is where capitalism reproduces itself (Lefebvre, 1992). It is from this capitalist colonization of the everyday that we come to normalize different conflicting narratives as they occur in our lives, and as such the everyday is where we can investigate the rhythms that people embody and habituate in the present.

Although Lefebvre provides a dynamic image of everyday life, it is Michel de Certeau's *The Practice of Everyday Life* (1974) that is widely considered one of the key works in the study of the everyday. De Certeau's theory recognized the conflicting forces at work in everyday life by distinguishing between strategies and tactics. Institutions and macro-level shaping of society by capitalism occurs through strategies, which people develop tactics for living in, through and against. Berlant believes that the problem with theory such as de Certeau's is that it no longer accurately accounts for the situation most people are living in today (2011, p. 8).

It must be noted that alongside the developments in cultural theory from Marx to Lefebvre, a correlated trajectory in anthropology beginning with Ruth Benedict and Bronisław Malinowski has been concerned with the act of studying culture itself through developing a methodology for accessing everyday life (Goldenweiser  Benedict, 1937; Malinowski, 1922). In order to develop academic validity to the claims and observations made by an anthropologist as to how life is lived somewhere, participant observation emerged as the best solution, requiring that "the researcher *participates* with people in commonplace situations and everyday life settings while observing and otherwise collecting information" (Jorgensen, 2015, p. 1). Instead of generalizing to speak of all people or all cultures, the early twentieth century with the work of both Franz Boas in the United States and Bronisław Malinowski in the United Kingdom and Europe set the standard for participant observation as the best way to discuss a specific culture's customs within their own terms, in hopes of breaking from the imperialistic past of anthropology (Jorgensen, 2015). While not overtly operating under the rubric of everyday life theory, participant observation's extended exposure and participation in everyday life has enabled scholars to understand how the everyday may be radically different around the world, but also allowed them to discover commonalities between these disparate processes. These methods have remained at the heart of anthropology, with ethnography standing for an active participation in one's hybrid position within and outside of a culture while exposing that which may be hidden to those who are too close (Jorgensen, 2015).

The colonial legacy that anthropology carries has left the political and ethical lines of ethnography contested, but has also encouraged scholars to consider cultures which may have been historically ignored or excluded from studies. One example of these new fields is the work of digital anthropologists, particularly those interested in the cultural and experiential functions of digital technology and the users of such (Hjorth  Pink, 2014; Postill  Pink, 2012; Hjorth  Cumiskey, 2018). Through focusing on aspects such as routine, movement, and sociality in digital spaces and via digital technology, digital anthropologists have attempted to address the significance of life online and the particularly apparent gap between the adoption of digital technology and its subsequent academic interrogation (Murthy, 2008). Within this domain of study, Larissa Hjorth's focus on the everyday, affect, and perceptions online through ethnography has been extremely impactful in illustrating how cultural analysis and anthropology can be extended to digital culture through new considerations of how deeply ingrained digital technology has become in everyday life (Hjorth  Pink, 2014). Additionally, Ralph Schroeder's work recognizes the need for social theory to account for the impact of digital technology, and proposes new theoretical conceptions of the everyday in doing so (Schroeder, 2018).

Another emerging approach that anthropological and cultural theory both took inspiration from and were influenced by is the affective turn (Clough, 2008). From the mid-1990s onwards, the Humanities and Social Sciences have seen a surge in attention to the affective which has been attending to "a dynamism immanent to bodily matter and matter generally – matter's capacity for self-organization in being in-formational" (Clough, 2008, p. 1). This can be described as theorizing and focusing on impact or qualitative changes, through focusing on a more dynamic understanding of the way things impact each other in the world. By developing new methods and theories for affect, scholars have been able to discuss ephemeral realities in a newly concrete and tangible way.

To position this turn within academia, the theoretical basis of the affective turn has roots in a long history of queer and feminist consideration of emotion (Ahmed, 2015, p. 9). In anthropology, emotion has emerged as a point of attention from the early work of Lutz  Abu-Lughod (1990), through more recent diagnoses of anxiety and fear as the experiential dimensions of globalization (cf. Bauman, 2018, Comaroff  Comaroff, 2003). Similarly, the work of experimental anthropologists such as Michael Taussig's *The Nervous System* and Kathleen Stewart's *Ordinary Affects* have challenged how ethnographic writing grapples with this dynamism and attends to the *something* which has force in this world but may not be easily apprehended in discourse (Taussig, 2016; Stewart, 2001). In the Humanities, this affective turn inspired work which focuses on affective structure and the formalization of it in producing cultural objects such as in Eugenie Brinkema's *The Forms of the Affects* and Sinnae Ngai's work on tone (Brinkema, 2014; Ngai, 2007). This consideration of affect's form is also what inspires Lauren Berlant's research, working to read genres of the affective present as affective structures (see Chapter 2), which provides a rigorous tool for approaching a variety of forms that may appear. This approach provides the basis for considering memes from within the everyday.

To address the dynamism of memes from the position of the everyday, Lauren Berlant's theory of the affective present returns to Lefebvre's approach to rhythms and patterns in opposition to de Certeau's approach. As discussed in the introduction, Berlant proposes that capitalism and other forces of today *disorganize* the everyday, leading to a sense of crisis ordinariness wherein the failure of past fantasies about having and building a life leaves us searching for new ones. Although this process of finding new genres of living and adapting to this crisis ordinariness in some ways addresses this disorganization and attempts to overcome it, it does not change the enduring need and desire for a good life (Berlant, 2011, p. 7-8). In this chaotic present, Berlant develops a formalist approach to affect which focuses on the dramas of adjustment and produced narratives of the new pressures of building a life. Within this, she focuses on the place of fantasy and the promises of the future via affective structures to explore how these train our intuition of the present in their consumption and replication. Using Berlant's criteria for genre to consider the meme, the next chapter will define genre and develop an argument for the meme as genre.

## II. On Genre - The Meme

In 1984, Frederic Jameson wrote his now celebrated essay ‘Postmodernism, or, the Cultural Logic of Late Capitalism’, in which he discussed a particular shift in both art and life that he named the ‘waning of affect’, stemming from his perception of increasingly depthless artwork (Jameson in Norton, 2018). The legacy of this argument remains, but Berlant conjures up a different and contrary approach to periodizing this contemporary present by leaning further into affect. Berlant's theory shifts the logic of Jameson to embrace a non-essentialist approach, focusing on the means through which we build intuitions and expectations of what being in this present moment is like. She mirrors Jameson's structure but denotes this period instead as “the waning of *genre*,” wherein the traditional realist genres of the past no longer hold under the current conditions of what it means to have and build a life (emphasis my own, Berlant, 2011, p. 6). Generally, genre can be considered “a classification of type or kind. . . [which] regulates the narrative process producing coherence and credibility through patterns of similarity and difference”, but Berlant stretches this term to encompass both lived and depicted processes which we experience and replicate (“Genre”, 2004). This chapter takes up this curious concept of genre in order to to disentangle the multiple meanings and uses that genre has and to develop an argument in support of the meme as a genre in itself. This process will incorporate Berlant's more recent work with Sianne Ngai on comedy to consider how the meme functions as a new genre of comedy in society. This chapter will conclude with an ethnographic example of a meme, to consider some of the implications of claiming the meme as a genre.

### Genre

The term genre is traditionally used in classifying literary and artistic objects, but Berlant transposes it onto the realm of affect theory to attend to the particular structures and conventions (aesthetic, temporal, and ways of living) which make the contemporary present apprehensible, tangible, and set expectations for future intuiting in the process. For example, Berlant considers such diverse genres in her work as “the situation, the episode, the interruption, the aside, the conversation, the travelogue, and the happening” to explore how the allocation of agency and structuring of temporality capture a particular perception of the world and its possibilities (2011, p. 5). Genre can thus be seen as the mediator which transforms the amorphous present into a mediated affect, sensed and under constant revision as each iteration of a genre further establishes how the forces may play out in this particular structure. This desire for new genres in the failure of the old ones show the “wearing out [of] the power of the good life's traditional fantasy bribe without wearing out the need for a good life” (Berlant, 2011, p. 7). By turning towards new genres and their new fantasies, we can consider how the contemporary present when framed in different ways "provide[s] an affective expectation of the experience of watching something unfold, whether it is in life or art" (Berlant, 2011, p. 6). These new genres emerge as new styles of managing "the narratives of what's going on and what seems possible or blocked in personal/collective life" (Berlant, 2011, p. 4). The question is, then: in what particular ways do memes as a genre set expectations and give shape to the present?

The need for new genres such as the meme is not simply an organic emergence at the hands of digital technology, but emerges from a particular historical present in contemporary society wherein our conditions are no longer represented by the genres of the past. The titular concept of Berlant's theory (cruel optimism) describes a structural relation in which. As discussed in the introduction, Berlant claims that the affective structure of cruel optimism as a sustained attachment to "something you desire [which] is actually an obstacle to your flourishing" (2011, p. 1) is reflective of the structure of our contemporary present as a whole, touching upon the anxiety surrounding failure in reciprocity from larger societal systems. Although the factors which compose our world such as institutions, the state, and our environment continue to fail and disappoint us in returning the work we put towards them, we can still recognize our strong, continued attachment to these factors and our continued investment. To handle this sense of extended crisis, we do what we have to; we all participate today in the improvisation of genres (simply put, we try out different ways of being and living), where the emergence of new genres today reflects our striving for stability and ways of being in a world that feels without any. These genres, which are made physical in a variety of lived actions and artistic depictions through replicating affective structures, have the potential to reconfigure our relations to our fantasies of the good life and what can be a good life for us.

It is important to specify what I mean when speaking of affective structure: although the *experience* of an affective structure may feel any variety of ways, the affective structure itself is one which can be deduced through formal consideration of subjectivity, fantasy, and form coming together in something (Berlant, 2011, p. 2). For example, optimism as an affective structure involves "a sustaining inclination to return to the scene of fantasy that enables you to expect that *this time*, nearness to *this* thing will help you or a world to become different in just the right way" (Berlant, 2011, p. 2). Another example can be seen in describing the genre of the situation: "a situation is a state of things in which *something* that will perhaps matter is unfolding amid the usual activity of life" (Berlant, 2011, p. 5). By establishing a form which can take on a variety of objects within it, affective structures and genres allow for specificity of identification without closing off subjectivity and situational specificity. In this case, the ability of the 'situation' to apply to both the situation comedy and the police interrogation ("We have a situation") emphasizes the absorptive nature of genre in Berlant's use of the term (2011, p. 5).

As such, paying attention to affective structures allows for a formally grounded consideration of how relating to the world is not a neutral act but is rather a process which involves the participation in and formation of new genres as they emerge. To pursue a cohesive definition of the meme, the next section will present Berlant's work on comedy to position the meme as a new genre emerging within comedy, and extending the author's work to the particular context of digital technology.

## Comedy

Berlant's account of comedy is useful for clarifying the complex and ambiguous place of comedy in life today. Berlant describes comedy as "always a pleasure-spectacle of form's self-violation" (Berlant and Ngai, 2017, p. 243). It involves the watching of something which somehow defies or violates itself, creating a spectacle from which to derive pleasure. This means that comedy itself is a particularly ambiguous and contextually determined form where deception and inverted logics provide pleasure. The comedic can involve questions of subjective perception, such as "what's living, what's mechanical, and who needs to know", which also establishes comedy as the scene of aesthetic, moral, and political judgement (Berlant Ngai, 2017, p. 234). At its core, comedy is based on intuitive intersubjectivity, whereby we test out what is shared and not shared in a more practical and discreet way than any traditional theory of comedy claiming to be objective could (Berlant  Ngai, 2017, p. 235-6). It is difficult to create a theory of comedy because we feel out what is comedic, with the formal lines of humor as experienced contingent upon the events and contexts of the world filtered through a subjective lens. As a result, we can view comedy as the pleasure of this reevaluation where we renegotiate what in the past had been perceived as harsh lines (Berlant  Ngai, 2017, p. 235).

One of Berlant and Ngai's assertions about comedy which I find particularly productive for this discussion is the suggestion that the 'comedification' of everyday life has placed comedy as an "overarching tone of late capitalist sociability, affecting how people self-consciously play as well as work together and the spaces where they do so (including Twitter, Facebook, Snapchat, Instagram, and YouTube)" (Berlant  Ngai, 2017, p. 237). As this is one of the only mentions of social media explicitly in Berlant's work, the inclusion of comedy in the discussion of memes as genre is not only warranted, but imperative to understanding the centrality of comedy to this realm of the everyday. Therefore, considering memes as a new genre of comedy within the digital shows how the contingency of the meme's humor is safeguarded by increasing the number of memes one may encounter, suffusing much of life with comedy. To start outlining some of these dynamics involved in the work of a meme and its relation to comedy, the rest of this chapter will use an ethnographic case study to explore the general structure of a meme and its relation to comedy as a whole.

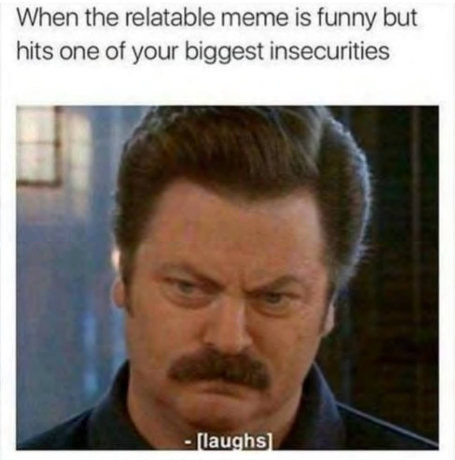

When the relatable meme is funny but hits one of your biggest insecurities
- [laughs]

*Figure 1* – Ron Swanson Meme (taken from observations, March 11th, 2019)

## Case Study: A Meme about Memes

> "The meme in question was sent to me on my first field site visit, during which I messaged a close group of my friends in a group chat where we frequently exchange memes. Asking for any memes that they had seen recently, many of the memes I was sent in this way were static images [2], as they were able to be saved from social media and directly sent to other users through personal messaging. This was the more traditional form of the meme in the past, although the development and changes in social media affordances have allowed for the inclusion of GIFs which loop a drama infinitely, or a video which adds an audio-temporal dimension to the drama at hand. The top text in this meme (see Fig. 1) reads "when the relatable meme is funny but hits one of your biggest insecurities," and the majority of the image has a picture of an angry or disappointed man (the character Ron Swanson from Parks  Recreation), with subtitle text on the bottom that indicates laughter in the original scene."

(observation, March 11th, 2019)

As can be seen from this ethnographic attempt at describing the meme, memes generally have an *ambiguous* appearance. While the meme itself is singular, it is never in isolation but rather defined in relation to the variety of memes which come to be absorbed within the genre of meme and further reproduced under this guise (i.e. Ron Swanson's character, "When you. . . " memes structure, using a cropped Netflix screenshot). Additionally, the combination of a framing phrase and a reaction image shows how a meme involves the work of context and content in their loosest interpretation. In a meme, context(s) (the form of a setting) and content(s) (the things which are held within this) can layer, mix, persuade, and combine to develop

[2] In order to disseminate across different platforms, many memes are either directly exchanged as images or are screenshotted and then disseminated through the system. Some of the memes I received directly as images are included in the conclusion.

new forms of content which are further contextualized to develop more comedy. In this meme, we see the context establishing a time but the place is not necessary. The language of this context establishes a disjunction between the subject and object in both the act of reacting to memes and in the subject who comes to 'view' it. By including an image of a man looking angry but with the subtitle of laughter, this disconnect is even more comical as we identify the meme's ability to capture what feels like a disjuncture of our experience through the content of this context.

Turning to the ambiguity of the context, we come to see how the instability of language emerges. Rather than this destabilizing of memes leading to a rupturing in meaning, meaning multiplies as the meme absorbs and comes to stand for an entire complex of objects, experiences, memories, dreams, and forces brought to life through one's ability to 'feel out' the scenario. Considering my own experience of this meme, I know my biggest insecurity, and this meme may suddenly lead me to consider not only this insecurity but its derivative pleasure as this meme transforms this vulnerability into an object of humor for myself. What's important to recognize is that the placeholder of 'my biggest insecurity' provides access to a deeper awareness of the self through considering at the time of viewing what one's biggest insecurity is; but this knowledge is only brought up for one's own interpretation and does not travel with the meme – making space for vulnerability without necessitating disclosure. As for what we bring to memes, this insertion of the self into a hollow context acted out by a relatable representation as in this meme is the work which invests the self into the digital while still cultivating some sort of further knowledge of the self in the process. Contexts come to build, multiply, and challenge meaning as the interface allows more engagement to layer normative modes of behavior in these different conflicting realms. Thus, we, too, can re-contextualize, decontextualize, and find ourselves in an array of situations made tangible by the meme. For example, the vignette featured in the introduction [3] highlights the rapid movement of the subject between different scenes within a single screen. But still, a meme only works when we are able to inhabit that representation, however flawed or misdirected the recognition in itself may be. This problem of (inter)subjectivity will be further explored in the next chapter [4].

Memes may be spoken of in passing as simply the source of a small chuckle, but it is this investment of the self into the meme through recognizing and living (in this case, playing out one's biggest insecurities) that we can begin to see the serious dimension of memes. The affective structure, the genre, or set of formal expectations which memes produce indicates how the pleasure and recognition of projecting the self (towards the meme and back to the self) attaches one to fantasies which may play out in the meme via this particular object. In this example (Figure 1), my biggest insecurity becomes generalized to part of the affective experience of viewing memes as a whole while I interact with the meme, where my own experiences cannot be excluded in considering these forms. This meme (see Figure 1) depicts the doubling effect of thinking about memes through a meme, showing how the process of viewing memes is made tangible in itself as an object of consideration by the meme. By making explicit the conflicting forces involved in my own extraction of humor from a joke about my insecurity, the affective structures of memes as a whole is made apparent. This is done by characterizing the scene of viewing memes as a framework for making these connections and renegotiating intersubjectivity. The meme draws in the viewer to invest themselves in the scene, allowing them to feel out new affective relations to their own experiences via this content by, in this case, engaging with their own insecurities in the past. Thus, it is the potential for new and different modes of inhabiting the self to emerge which furthers the self's quest for development by viewing and interacting with a meme. The collective body of memes as this growing space for new and complex ways of considering how we relate to the world and what it means to have and to build a life at large, is as inseparable from the meme as its status is contingent within the context of memes macroscopically.

The meme considered within its inextricably interlinked relation to the entire body of memes provides the space for developing new, less rigid

[3] See end of the Introduction

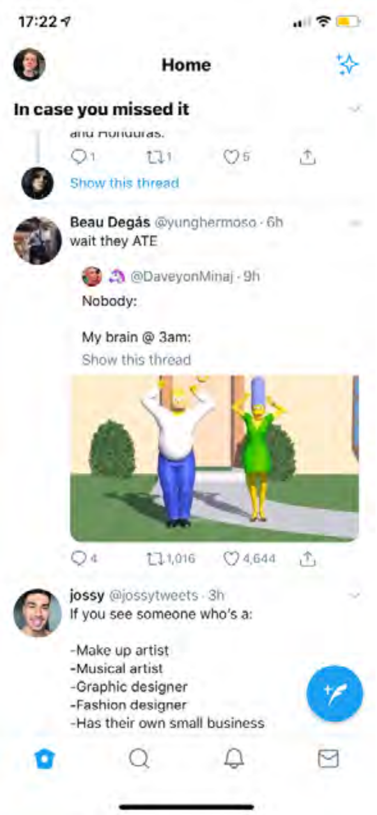
[4] See Chapter 3

forms of relationality which the world increasingly requires of us in the hopes of deriving pleasure from recognition itself. The viewing of memes is a practice of relief (as I will further elaborate below), but it is, nonetheless, not devoid of or divided from feeling out the practical necessities of the present. It involves the assessment and judgement of different means of managing a life, and involves the celebration of both success and failure as it is reproduced beyond our own precarious inconsistency. Thinking of internet memes as a subgenre of comedy provides a new lexicon for attending to the affective work in how memes come to mediate an otherwise confusing and overwhelming world without collapsing into another discussion of comedy. In categorizing and setting expectations of what viewing memes may be like, such that we expect and thus demand comedy from the meme, our own set expectations are troubled in the process. The rest of this research will take up the aspects of subjectivity (Chapter 3) and intuition (Chapter 4), which are both important in considering how the meme as interlocutor between the affective present and ourselves is (re)negotiated from both top-down and bottom-up means of control and engagement.

## III. Making Me(me)s – Subjectivity, Composure, and Self-Connection

> "I picked up my phone, saw some notifications of a new text, some news, and that app I kept forgetting to delete. I ignored them and unlocked my home screen, pushing upward and away what at the time was a richly-saturated surrealist body painting, painted on and by Dain Yoon's face with a distorting effect which makes her ripple in front of a red background. Tapping on the twitter icon and entering into the light blue and white interface, returning instantly to a post which I had last seen earlier in the day, I find myself back to a space which is familiar to me.
>
> A quick press of the top bar, and I find myself whisked abruptly past an ever-growing number of tweets to whatever post happens to have been the most recently posted to me. Within an instant, I'm confronted by an acquaintance from New York announcing that he'd just finished some important exams, a video of FKA Twigs' makeup that was retweeted by someone I don't remember which someone posted saying how she has no flaws, and a post from a drag queen which was liked by someone I follow named 'b' with a video of someone putting on a clown costume commenting on how she would be wearing that outfit tomorrow ("me tomorrow"). I start to go through them one by one, working my way down to the line break which starts a new flow of posts under the heading of 'In case you missed it'. (Usually, it's not in case but rather what I did miss at some point, what have I missed?)
>
> I stumble across another post on my feed (see Figure 2) and I look at it. First, I see someone I follow who retweeted the post and captioned it as, "wait they ATE," which I look up on Urban Dictionary to see means that 'they' performed something well. I open up the post (see Fig. 3), intrigued by a high-speed video of an animated sequence of people who have been personalized to appear to be Homer and Marge Simpson, dancing for 40 seconds to a recently viralized portion of Nicki Minaj's Roman Holiday that has been edited over many videos. The original post says, "Nobody: | My brain at 3am:," with a video completing this rebuttal to no one's action. They perform even faster than the edited song, showing that this sequence must have been performed in some context, and then found and added to this. I laugh, because between the speed, the seemingly illogical organization, and yet still skilled performance, something in the meme really does capture what it feels like my brain is doing at 3am. I don't watch a lot of the Simpsons anymore, but still, I relate."

(observations, May 4th, 2019)



In case you missed it

and Honduras.

Show this thread

Beau Degás @yunghermoso · 6h
wait they ATE

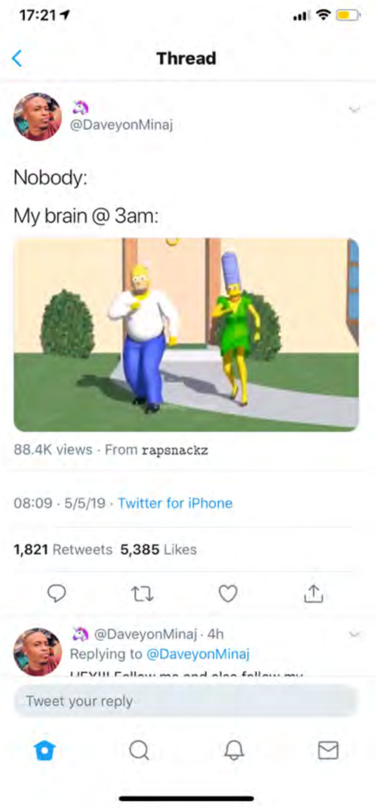
@DaveyonMinaj · 9h
Nobody:

My brain @ 3am:
Show this thread

jossy @jossytweets · 3h
If you see someone who's a:

-Make up artist
-Musical artist
-Graphic designer
-Fashion designer
-Has their own small business

*Figure 2* – "wait they ATE"  (observation, May 4th, 2019)



@DaveyonMinaj

Nobody:

My brain @ 3am:

88.4K views · From rapsnackz

08:09 · 5/5/19 · Twitter for iPhone

1,821 Retweets 5,385 Likes

@DaveyonMinaj · 4h
Replying to @DaveyonMinaj
HEY!!! Follow me and also follow my

*Figure 3*– Nobody | My Brain @ 3am  (observation, May 4th, 2019)

A meme, like the one described in this segment of fieldwork, is like an inside joke. That is, to see a meme is like being *inside* of the joke, or to complete the circuit of the meme and to engage a moment of recognition through inhabiting and feeling out a space, time, or moment such as my brain being represented by this video. It is not the fact that I know who the Simpsons are, or because of a particular affinity with Nicki Minaj, but rather the combination of disparate objects into a context which I am attached to (these things allegedly representing my brain) that brings these objects closer in proximity to my life and compels me to relate to them. Most jokes, as Ngai and Berlant (2017) argue, work to draw on ideas of "what it means when we say 'us,'" as we navigate how we relate to others and particularly the many identities we may be taking on (p. 235). Our sense of self and our knowledge also affects our sense of literacy, a teaching by the world of meanings and messages from which one can extend to digital spaces and make further connections.

There are few requirements for how a meme must look, be, or express, but memes always involve a 'me'. To consider the meme without the place of the viewing subject would be to assume that the meme is neutral, that the meme does not affect me but merely represents me (or something to which in some way I relate – such as the video above). The subject is the one who gives life to a meme, not only in activating its logic but also in the potential for its distribution and further uptake as a carrier, standing between the meme and one's own network. Memes involve the democratic navigation of what we like, and what it's like to be us – but to see the memes without considering how subjectivity may be shaped by these objects is to disregard how the creation of these bites of bytes involves the materializing of this affective present which I am attempting to tap into.

Thus, this chapter will mobilize Berlant's conceptual framework to understand subjectivity - or the relation to and sense of one's self - within the meme. This will begin with an overview of subjectivity within Berlant's work and then discuss the modes of subjectivity which memes reinforce through comedy, drawn from this ethnographic fragment above. Finally, this section will consider how memes impose and reinforce new modes of subjectivity which blur the lines between self-continuity, self-expansion, and self-interruption in the cultivation of subjectivity online.

## Berlant, Lefebvre, and Sedgwick on Subjectivity

Firstly, a consideration of the present is inseparable from a consideration of subjectivity, because subjectivity is the condition of being a person as well as the process of becoming one within this present ("Subjectivity," 2004). As this research focuses on the present moment which Berlant describes as disorganized, unstable, and particularly uneasy, subjectivity stands as the tool through which we simultaneously navigate our sense of self and the world. Ergo, our own subjectivity shapes how we see the world, while this subjectivity is trained and reshaped through the work of genres by setting expectations to reconfigure this sense of self. For Lefebvre, it is society's normative conventions imposed on us which make us see ourselves as individuals (and this individuality as a reality of existence), while what we claim to be this individual subject – what one could call a personality – does not simply reveal or express an innate inner essence. Instead, it is our actions of individuality which produce the potential for habits, securing physical safety while allowing for intellectual freedom, a securing of life's continuity to allow for life's betterment in the now (Berlant, 2014, p. 63). One's personality forms a history brought forward, and imposed unto the now as a framework through which to see the world. Although we have this history, the present is full of potential and such a sense of agency that we feel as though we could truly act in any direction we may choose. To resist this nihilistic sense of potential and infinite possibility, we lean into this personality as a sense of grounding and come to conceptualize our positioning as not the result of society's active shaping but simply as who we are – our sense of an innate self. To unpack the shaping of such subjectivity by social forces, Berlant draws on Lefebvre's model of *dressage* to speak of subjectivity as a process whereby we train ourselves and are trained to behave in particular accepted and expected ways, not through active cognition of such a process. Rather, by emphasizing the importance of social practices and rhythms Lefebvre turns our attention to rhythms, both biological and imposed, to consider how these rhythms comprise and produce subjectivity *in itself* (Lefebvre in Berlant, 2014, p. 198). Berlant and Lefebvre provide a perspective that is fruitful in attending to patterns of meme consumption and construction, to consider the force of these processes as self-making practices.

Lefebvre is not the only source of inspiration for Berlant's investigation into subjectivity within the contemporary present. Drawing on Eve Sedgwick's work into alternative understandings of desire and the self, Berlant departs from the dialectically traditional formation of neoliberalism's force as imposed on the subject yet subjectively autonomous, and instead looks for something much messier (Berlant, 2014, p. 123-4). Sedgwick makes space for us to think through "a practice of meticulous curiosity" as Berlant calls it, for the potential of a reparative reading of agency and desire that recognizes where those practices of therapy and relief are worthy of study, too (2014, p. 123-122). Although there's been great work done trying to highlight memes as a serious matter, this research follows Sedgwick's mission and instead focuses on the joy and pleasure that can arise from people thinking and living together in digital space. As much as this chapter criticizes the way in which memes shape and alter subjectivity, it is also possible to embrace the reparative potential of tracking attachments in hoping to place them "back into play and into pleasure, into knowledge, into worlds" (Berlant, 2014, p. 123). Between Sedgwick's acknowledgement of the complexity of attachments, and Lefebvre's dressage to explore the impact of practical rhythms in shaping these attachments, Berlant's approach to subjectivity provides a way to consider the shaping work of memes upon the subject and the subject's agency in concurrently shaping the meme as well.

## Composure and Self-Interruption

It is fair to claim that not everything we do in this world is in line with the idea of a self and a personality which we have or hope to live so far. Personality provides comfort, but it also sets parameters for how we must behave based on the history which we experienced and how we proceed from this. When someone acts 'out of character', what character is it that they are required to live within? Can being someone else - or being just a little less trapped in the self you have been told you are - even for a moment provide space for relief?

The conceptual bridge between a subject's personality and this subjective escape from the self which I will henceforth call impersonality is what Andrew Phillips calls composure, or the relationship one has in relating to their own mind (Phillips in Berlant, 2014, p. 144). Berlant, in her own words, calls composure "the safeguard of fantasy," as it allows you to "set the scene for your entrance and make the world come to you when you want it" (Berlant, 2014, p. 144) acting as a buffer to the world. Through composure, the mind can make space for feeling out a different sense of self that is not based in the practical continuity of the world and self-continuity of action, but rather operating perpendicularly to this. Composure makes enrichment possible through an understanding of the mind's enrichment as something worth pursuing, seeing how acts of relief and enjoyment enrich our lives while not necessarily aiding in their continuity.

Composure allows for the interruption of the self, emerging in situations of perceived safety that gives the mind room to think of itself and the world in new ways; composure protects the self while also allowing for new forms of recognition that are not caught up in the ongoing logic of things (Berlant, 2014, p. 144). Berlant proposes a model which claims that "the body and a life are not only projects, but also sites of episodic intermission from personality, the burden of whose reproduction is part of the drag of practical sovereignty, of the obligation to be reliable" (2014, p. 116). These intermissions can take form in activities such as eating, having sex, and looking at memes like in the ethnographic fragment above. This involves "inhabiting agency differently in small vacations from the will itself," drawing on many kinds of self-understanding and often works towards "making a less bad experience" (Berlant, 2014, p. 116; p. 117). For example, in the case of eating, Berlant describes this as a mode of self-extension whereby the subject feels out something about the moment (a taste, flavor, sensation), to cultivate "a sense of well-being that spreads out for a moment, not a projection toward the future" (Berlant, 2014, p. 117). While composure enables a variety of modes for subjectivity which Berlant explores, such as self-interruption, self-extension, and self-suspension, the meme requires new ways of thinking about life-bettering practices online.

Turning to the ethnographic fragment at the beginning of this chapter, the internet emerges as a space in which one faces the pressures of socially required personality imposed upon an intrinsically impersonal interface. For example, technology such as my iPhone arrived to me in default factory settings, but I have continually made my phone and the content of my applications into my own image since. This personalizing of impersonal technology has the double-edged effect of also producing a digital body in the form of a profile - one which the viewer inhabits first handedly while creating and producing a public facing portion of a profile such as those I repeatedly stumbled across in my research. On social media, through the acts of scrolling, following, liking, blocking, and messaging with others, this pushing and pulling of content around the user makes space for their connection to a gargantuan network of connections, while also elaborating on their own self through their connection with others. However, this separation between physical self-continuity and digital self-continuity (which is not in itself compulsory in the same ways) is what has allowed for the inspirational and optimistic scene of the online. This mix of personalities and impersonality is fundamental to the internet as a technology which is being made personal. With their ambiguous appearance, memes further carry this oscillating flow.

Where the sense of crisis ordinariness that Berlant established can leave life feeling "more like desperate doggy paddling than a magnificent swim into the horizon," the internet and social media have provided the means by which the social can be accessed *impersonally* and texitítremotely while requiring the creation of a profile for one's own participation (Berlant, 2014, p. 117). Therefore, although the digital is separate from our embodied lives, there is still an alternative space for a different sort of continuity if one chooses. In the case of the meme, we are neither forced nor required to participate in a particular way but rather are confronted with memes based on our networks and are subsequently able to make judgements on a piece of content and act accordingly. In this process, we are taking in new ways of self-understanding and developing new conceptions of how we fit within this larger virtual community made legible and logical to us in many small chunks which we can find and track. Such as in the vignette at the beginning of this chapter, it is the formal and affective structure which provokes these connections, as it is the structuring of a relation mediated by a relation which compels the subject to turn towards these connections more than to the content of the meme. What the meme does is present a relational structure, and allow the subject to gauge their own connection or disconnection with this relationship.

Thus, I propose that the meme provides a timeless (yet contingent) scene to which the self can return and experience a mode of *self-connection*. Building from self-extension, self-connection involves an extension within a moment rather than towards a future, but one whose main focus involves the navigating of intersubjectivity on many levels. Recalling my argument in Chapter 2 that memes constitute a genre related to comedy, I turn here to the intersubjective function of comedy to consider how the meme generates self-connection, and to further explain this case (Figure 2).

### Comedy and Intersubjectivity

Comedy, as Berlant and Ngai suggest, relates heavily to aesthetic judgement, as they both remind us of "forms of intersubjectivity we usually don't think about but that we rediscover as presupposed by our very compulsions to make jokes and judgements in the first place" (2017, p. 235). The foreshortening of comedy is made even closer through social media at the touch of a finger, and the curation of posts into a news feed creates the endless stream of content expansively relating it other content, leading to this layering of contexts developed in Chapter 2. As memes involve making claims about and judgements of these relations, they extend and grow through further layering of new meanings, not only as the self cultivates new and interesting ideas but as you see others seeing it, when they post it, and who else may have liked it. For example, the chronological distribution of posts by the users one follows into a personalized feed can happen to arrange content into new formations that the meme could not predict. Not only can memes such as the one above combine a confusing combination of objects (in this case Nicki Minaj, an amateur animation of the Simpsons, and a classic "Nobody | Me" meme form), but the recontextualizing of adding the comment "wait they ATE" by another user redirects the attention from the relation of the content to your own experience. This then is considered in relation to their technical skill, which could be extending one's self-awareness through a consideration of gracefulness in performing this dance. Through the rabbit hole of considerations that multiply from ambiguity, the diverse ways in which a meme can be redistributed and remixed can draw on and extend one's own existing sense of self-connection to the networks online in the process.

The meme provides the pleasure of recognizing new forms of intersubjectivity and laughing at what arises. In doing so, the subject embraces an attitude where any form of knowledge that which can be logically made apparent (even if the logic is visual, incomplete and shaky at best) can provide a punchline of closure to an otherwise constantly unfolding and unpunctuated present. The product of a meme is pleasure: often in the form of laughter from simplifying the world, but most broadly from recognition. The ability to cultivate an alternative self(s), and these small moments of affinity found through memes may motivate us to carry them with us as the digital re-enters our navigation of the real. The meme also reveals through its replication that others feel something like 'my experience', such as in Figure 3 which shows that I can feel less alone in being awake and thinking at 3am. Even considering how I have not had this feeling for a long time, the meme still takes me back there, and in that moment, I find a newly mediated space of affinity. My anxiety is no longer a violent force working against me, but through composure transforms into the source for my derivative pleasure. In some ways, memes draw on the already established genres of knowledge dissemination amongst visual culture to consider the kinds of pleasure we can generate from recognition through the impersonal, using the internet as a means for connecting to many others and drawing out new selves in the process.

### Recognition as Form

While memes contain absurd and entertaining content, this content only matters in appealing to the subjective taste of the user and their potential literacies. The meme as genre creates a space of affinity, wherein composure allows for the pleasure of recognition, with the potential for various forms of further recognition. Memes, as well as comedy, are incredibly sensitive to context, so it is this constant (re)negotiation of intersubjectivity that generates comedic pleasure while exposing and developing new ways of relating and connecting in the process. We focus on the meme as a source of pleasure, but the pleasure comes from recognizing your own self, recognizing your relation to the poster's self, and considering the wider connection to a topic, an opinion, or even the world (or how only in the meme an idea may *be made* possible). Memes can break through our stance of impersonality to collapse the space between the personal and impersonal, as we are both absorbed into these "me"s while absorbing further knowledge about myself in the process.

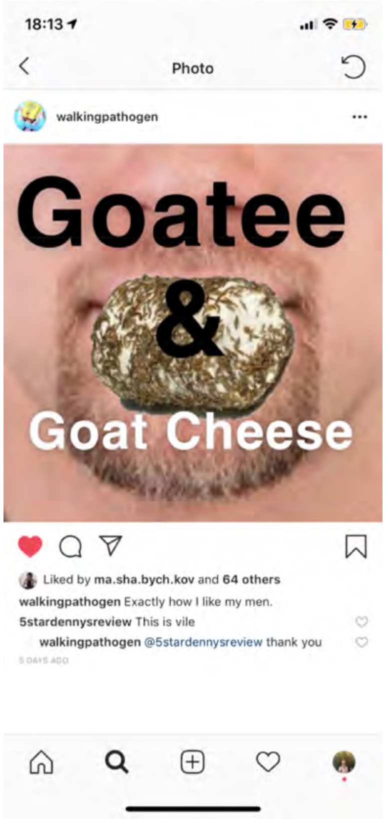
In terms of the expectations of the genre established by our interaction with the meme, the desire to find pleasure within the meme leads to what ultimately has been an impersonalization of the personal. Therefore, representation of the past is no longer required as I can project myself into a variety of contexts and find recognition in objects as diverse as in Figure 2, where I come to recognize my brain as a crudely animated pastiche of cartoon nostalgia converted to artistic performance. It is the development of self-connection and openness to comedic relations which builds the momentum of meme consumption, and ultimately forms a constantly-in-flux domain of relations.

Thus, memes are like inside jokes because not only are we positioned inside the joke, but the joke places the user inside of their own reality. One may also find that this new meme has something familiar which brings the pleasure of a past experience into the present through this new piece in unanticipated directions, allowing one to appreciate the stability that perfect reproduction allows for as its connections can move in unexpected and new ways. Encountering memes can be likened to warming your cold hands on the residual heat from the flurry of life taking place online, basking in the errors we would normally ignore or disregard while considering how taking things seriously only sets up future disappointments. To create a meme is to make a claim to one's way of adapting to the world, allowing others to momentarily assume this position by animating their own experience for comedic effect. The challenge, to be taken up in the final chapter, is to consider how this new mode of subjectivity requires new ways of paying attention.

## IV. Feeling Out and Through – Ideology and (Virtual) Intuition

> "Starting my very first field visit on Instagram, I am unsure as to whether I should start with my feed as it is in front of me, or if I should start on a meme page instead. How would I be able to discern if something is or is not a meme? I decided to start with the meme page, @walkingpathogen, since then I could be able to get an idea and make my own discretional choices of what to include and what not to include. The first meme I saw on this page (see Figure 4), however, already brings these concerns back to my attention. The image shows a close up picture of a man's goatee, with some golden foil object superimposed over his mouth. On top of this image is the text, "Goatee Goat Cheese," with the first half in black text and the second half in white text. The typed caption of the post says: "Exactly how I like my men." While I know that it is a meme for the fact that this meme account posted it, or at least I would assume this is the case, it doesn't feel to me like a meme. It's about goatees and goat cheese, and I'm guessing it's making fun of these objects as certain aspects (in some way) of who people are, but if I were asked what this image is, I would probably call it a meme from the context more than from the image itself. I'm not sure if it's because of my own taste in memes, or because I'm not

> accustomed to this one, but I wonder whose meme is it, then?”
>
> (observations, March 11th, 2019)



Liked by ma.sha.bych.kov and 64 others

walkingpathogen Exactly how I like my men.

5stardennysreview This is vile

walkingpathogen @5stardennysreview thank you

5 DAYS AGO

*Figure 4* – Goatee  Goat Cheese  (observation, March 11th, 2019)

Subjectivity speaks to the place of the individual within larger social dynamics, but this chapter considers the training of intuition as the mediator between subjective experience and collective biography (Berlant, 2014, p. 53). Building from the account of subjectivity developed in the previous chapter, we can consider how the mode of self-connection, which is enabled through the meme, draws on a form of *virtual* intuition to access and respond to the complex, performative space of the digital. To access this space requires new modes of being a subject (i.e. “*I’m* baby”) and new ways of paying attention to the world through this subjecthood (i.e. “*That’s* a Big Mood”). As the vignette above shows, it is in the pursuit of drawing meaning and pleasure from the meme where intuition is stretched and warped to encompass a variety of possible meanings that are made available and possible by *virtual* intuition. First, this chapter will discuss ideology in relation to intuition in order to ground this discussion of intuition in the collective negotiation of the present which forms collective ideologies. Then, I will propose this concept of virtual intuition and discuss its particularities in relation to my fieldwork. Finally, this chapter will end by considering how comedy relates to self-connection and virtual intuition, speaking to the effects of the genre as a whole in shaping the contemporary present.

## Defining Ideology and Intuition

Ideology, as defined by Louis Althusser, is the, “imaginary relationship of individuals to their real conditions of existence” (Althusser in Felluga, 2015). This definition draws on Jacques Lacan’s use of the term, *imaginary*, to identify ideology’s grounding in the similarities between people’s perceptions which enable us to give meaning and make sense of the world (Felluga, 2015). There can be many different kinds of ideologies, such as religious, political, and moral, which operate in many different ways, but ideology is ultimately a material condition which cannot be separated from the apparatuses of society which reinforce particular ideologies. Interpellation, or being hailed by an ideology, is a constant process of indoctrination which converts concrete individuals into subjects - a process that is already underway but is still susceptible to radical change. Ideology, therefore, cannot be separated from its practices, as these practices are “our performance of our relation to others and to social institutions that continually instantiates us as subjects” (Felluga, 2015). By considering genres as material objects which mediate the affective present, Berlant suggests that we must turn to the topic of intuition to address ideological concerns (Berlant, 2014, p. 53). Thus, to consider genre as the normative repetition and development of a collective relation to the world which builds expectation for the future, is to make space for the importance of intuition in genre as a tool for renegotiating ideology.

Berlant defines intuition as “the process of dynamic sensual data-gathering through which affect takes shape in forms whose job it is to make reliable sense of life” (2014, p. 53). This is not to say that intuition is simply the process by which people interpret and respond to the world in time, but rather sets up the claim that visceral responses are trained as we gather genres, set expectations, and make tangible the emerging event (Berlant, 2014, p. 52). As Berlant aptly describes, “intuition is where affect meets history, in all of its chaos, normative ideology, and embodied practices of discipline and invention” (2014, p. 52). Although ideology is important to consider in this context, it is intuition which provides an avenue for considering what feels ‘normal’, what is required of us both macroscopically and within the everyday, and how this may be changing.

## Virtual Intuition

Intuition provides a way of adjusting to the world and, as genres solidify into a habitual form, a way of managing life according to the assumptions we hold of how to manage a life (Berlant, 2014, p. 52). That being said, I believe Berlant’s concept of intuition can be taken further, as interacting with memes demonstrates what I call a form of *virtual* intuition. Because memes can solely be accessed via digital technology which limits the depth of our sensual engagement, this virtual intuition emerges as not only a way of trying to catch up to the present, but also to sense a world beyond that which appears before us.

Memes provide a scene, wherein the promises that cluster around a meme can be reproduced and re-encountered by seeing something familiar in new contexts, allowing these elements to gain new meaning through conversation with other users and memes simultaneously. Therefore, participating in memes requires a sense of virtual intuition – to savor the humor, to search for more comedy, to continue in this collective production of humor which ideally will make everyone (or those you would group in with ‘us’) laugh. Virtual intuition, then, is like suspending disbelief for the sake of seeing what comes out the other side, to allow for the virtual relations which have not mattered or have not yet emerged to exist and be shared. This is not to say that one will believe anything a meme shows, but that the meme’s job is to provide enough of a form around disparate content to hint at a relation which virtual intuition identifies and measures against the world. As memes are each made by someone with a particular subjectivity, virtual intuition also involves the consideration of where relations can be left open, and the possibility to revel in humor that, if it existed outside of the internet, may not actually be considered as all that funny.

Ironically enough, as the *content* of memes crosses lines and makes new comedic connections, the structure of scrolling through memes involves encountering many memes to absorb new forms of content constantly, sensing out how things are happening now. If a meme is not funny, then it may be too ambiguous or not provide enough of a reaction to the context. Yet through seeing more examples of the meme elements in varying contexts, these elements may become freed up enough for them to make sense and to provide pleasure for the user, such as Roman Holiday’s necessitated accompaniment of ‘fast motion’ videos being separated from its place in Minaj’s discography like in Figure 2 [5]. Virtual intuition also involves negotiating new meme norms, which may emerge alternatively to the real world but with their own standardized logic. Whatever the originally descriptive situation might have been (I do not claim to have ever seen the origins of this trend in this research), virtual intuition has also made space for similar fast motion videos to be shared which are then contextualized as lacking or missing the song’s presence. This shows how the ephemeral emergence of a meme form can solidify into a trend which trains virtual intuition to make connections based on the funny memes one has seen *towards* content one encounters. Therefore, we come to sense the contours of digital content for its possible connections through our virtual intuition which is developed by our interaction with memes and then expected as projected forward. The content then returns to and fits into our world, left open for connections yet solidifying into a trend which may become viral, or even reflective of the year it circulated. The meme, through its perceived accuracy and ability to describe (or what we intuit to be an inner logic from the chaos), can transform even the most absurd or ambiguous content into a stable, comedic trope solely through its repeated and frequent repetition and remixing. While formal meme norms may solidify through replication, these norms cannot prevent the content from grasping for or attaching to alternate possibilities, reimagined through this virtual intuition.

[5] See Chapter 3 for this argument.

We can also see how virtual intuition is important for accessing and navigating the claims underlying a meme. We can do this by considering the experience of encountering a political meme from an ideology that the viewer does not subscribe to, and seeing where the joke may be lost. This demonstrates how the unpredictability of memes makes the pleasure of identification with the joke just as contingent as the meme itself. Although memes have the potential to recognize the viewer, it is up to the viewer to intuit their relation to this recognition, and particularly to sense out how the location of this affect and its network effects the meme. This offensive potential of the meme, if the comedy does not foster a sense of intersubjectivity, is where the subject can experience the feeling of being the butt of the joke made by an excluding 'them'.

To investigate this sense of virtual intuition, I devoted time to exploring Far-right or Alt-right memes on Instagram through searching for 'right memes', which generated multiple pages. When I tried to look at these memes, I wrote this:
*"As a liberal/left person, I find that these memes annoy me because there is something so humble or inviting to the meme, and to see a meme from a view one does not support either makes you feel like the butt of the joke or like something about it is too far from the reality we live in, that the story is missing that part that makes justice possible. Funny stuff doesn't need to make sense, but unfunny stuff makes what feels like the 'wrong sense'."* (observations, May 4th, 2019)

While Althusser said that interpellation happens to us without even being able to resist it, I see virtual intuition as a more active form of intuiting the world, as the digital provides the space for a buffer of composure and the ability for a user to control their own experience, actively affecting their future experiences through judging in the now. In the case of my fieldwork, I found a collection of memes that seemed similar to those I had seen before, but the content had been shifted while appropriating a form for purposes that, to me, were not accurate. As a result, I do not follow or like these sorts of posts, but it made me think about this process of seeing memes that somehow resist my desire implicates a further network of users and relations that I do not consider. When a meme 'works', we find the expectation that if one thinks in the 'right way' (or right in relation to the creator of the meme), a meme *will* 'work' which requires considering what sorts of accounts one aligns with and the sorts of communities one wants to connect to. As much as an escape can provide an unexpected source of humor, memes produce a product of intuition; where one's ability to virtually sense out a possible meme or to create a meme can somehow still make someone feel recognized, and somehow confirm their place in this world. On the other hand, memes also provide a humble way of indoctrinating someone into a particular mode of intuition, in ways we may not be fully able to recognize in the moment.

## Feeling Funny: Comedy and Intuition

Compared to the refinement and training that intuition usually operates through, comedy is about surprise, the unexpected, the ability of a genre to collapse the space between objects, to allow them to play out in new and surprising ways, and to ultimately develop joy or laughter from what feels like a freedom that surpasses the limitations of the world causing a sort of disturbance (Berlant Ngai, 2017). It is difficult, then, to consider how we can be both expected and pressured to perform comedy in a growing number of zones of life while also holding that comedy itself involves the pleasure of sensitive and contingent relations being disturbed in different ways. This, I argue, is the importance of virtual intuition; in being attuned to the digital flows, as well as the real world experiences, one is able to combine both disclosure of experience with comedic interpretation, projecting a sense of self that is comedic from their participation in the network. As comedy upends assumptions to challenge form, virtual intuition is the process whereby more flexible relations to the world further protect the subject from its closeness, while feeling out these assumptions and identifying where comedy may disturb (hopefully) without rupture. Comedy is the consolation prize for adapting intuition to the changing world falling apart, because if one can stage the scene in just the right light, there will be pleasure derived from it. In social media, especially, these applications arrange the posts in any variety of ways, but virtual intuition looks for the patterns in the roll of dice, unintentionally creating new connections and breaking old ones.

It is the perfect multiplicity of the internet - in its accessibility, its (assumed) representation, and access - that we find a sudden appreciation for the flaws and errors that at once hurt and re-enact failure as a performance. We can fight, love, confess, and proclaim to the public in unprecedented ways and the worst that can happen is, ultimately, having one's view challenged for deviating from the norms of others. Instead of developing new stories and new characters to seek representation through, memes take the world as a means of expressing itself. They draw on our own lives which frame our experience of the world for inspiration as our own precarity in the contemporary present is funny enough. In hopes of connecting through a shared sense of pleasure from self-connection, the expectation developed around memes is that these digital objects *will* be funny, as much so as we can 'get it' or 'like it', if getting the joke feels worth it. This is where the sociality of the internet is incredibly important and virtual intuition necessitated, as the networks and relations one develops play a role in how we expect or understand certain content to be funnier than other content.

I'm not saying that memes change the way we see the world, but I do believe that they have changed the way that we relate to and affectively encounter the world around us. Memes paint the world not in events and processes, but by recognizing how the crisis ordinariness of life today is already funny enough. Thus, connecting to others rather than to pre-produced genres of the past does not make promises it cannot keep. As long as one is able to feel out this scene, the worst that can happen is that on the other side of the meme, things are still the same but with a new way of possibly and potentially seeing it. The prize, however, is the affirmation of one's mode of intuition itself through widespread normalization.

## Conclusion: Between the World and Me(me)

As much as theory can produce a deeper understanding and appreciation for the world, it can also warp and distort reality in that same pursuit. As comedy emerges like an unexpected pleasure of failures and faults, theory can work to retrospectively speak to comedy. Even then, to address and not totalize the subjectivity at hand is difficult. By speaking to the affective structure rather than subjective experience, this research has demonstrated that we can recognize the facets and dynamics at hand in memes without forcing memes to behave in a specific manner. Instead, embracing the immanence of affect, this theory I have extended from Berlant provides the starting point for further consideration of the affective contours of memes and the digital in shaping our present as we continue to shape them in turn.

To conclude, I would like to allow - in the canon of cultural analysis and style of its founder, Mieke Bal - for the memes to, "speak back" to my own exhibiting of their form, by providing a series of memes drawn from my fieldwork which have contributed to the development of this theory (Bal, 1999). As much as I consider self-connection and virtual intuition useful for explicating the meme, there is much more that can be said about memes, even about the ones featured here. Instead, I encourage those who read this work to consider how it may speak to their own experience. I hope that it may help some to return to their disciplinary approaches and incorporate an affective approach to memes in their study, as considerations of affect make tangible that which underlines our experience of a range of memes. Ultimately, the meme is a tool for considering the overlooked, attuning the self more flexibly to not only the world, but to the sorts of perceptions that are being circulated at the time. If we think of the meme as an affective mediator between the world and me, we can finally see their importance within the digital, the everyday, and in this contemporary present as grassroots interpretation and critique through comedic judgment. What is a meme without a *me*, after all?

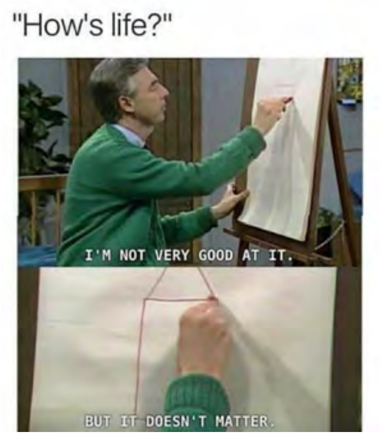


*Figure 5* – “How’s Life?” (observation, March 11th, 2019)

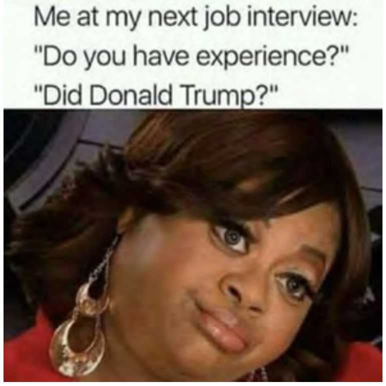
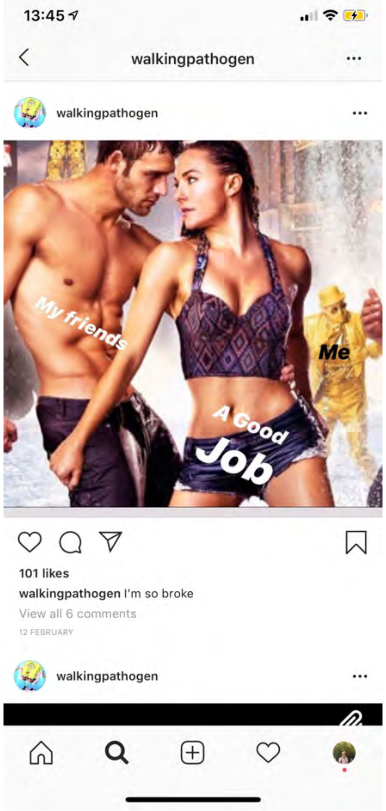


*Figure 6* – “Me at my next job interview” (observation, March 11th, 2019)



13:45

walkingpathogen

walkingpathogen

101 likes

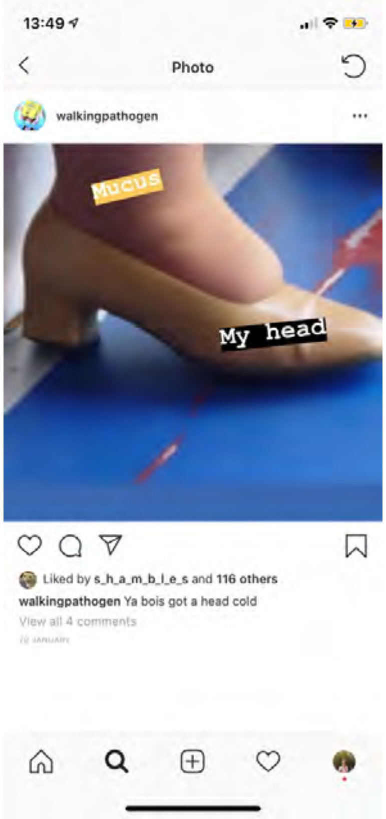
walkingpathogen I'm so broke

View all 6 comments

12 FEBRUARY

walkingpathogen

*Figure 7* – “I’m so broke” (observation, March 11th, 2019)



13:49

Photo

walkingpathogen

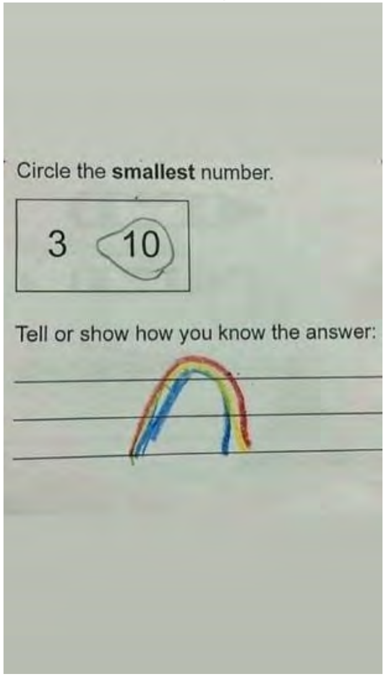
Liked by s_h_a_m_b_l_e_s and 116 others

walkingpathogen Ya bois got a head cold

View all 4 comments

12 JANUARY

*Figure 8* – head cold (observation, March 11th, 2019)



Circle the **smallest** number.

3 10

Tell or show how you know the answer:

*Figure 9* – smallest number (observation, March 11th, 2019)

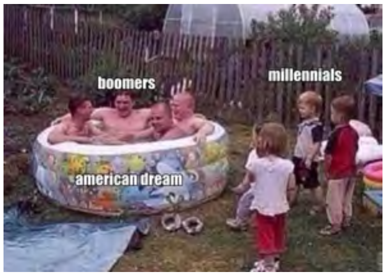


*Figure 10* – American dream (observation, March 11th, 2019)

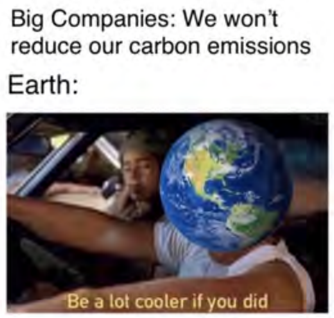

Big Companies: We won't reduce our carbon emissions
Earth:
Be a lot cooler if you did

*Figure 11* – Be a lot cooler if you did (observation, March 11th, 2019)

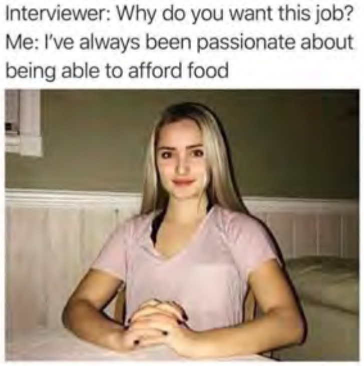

Interviewer: Why do you want this job?
Me: I've always been passionate about being able to afford food

*Figure 12* – Why do you want this job? (observation, March 11th, 2019)

### Works Cited

Adams, T. E., Ellis, C., Jones, S. H. (2017). Autoethnography. *The International Encyclopedia of Communication Research Methods*, 1-11. doi:10.1002/9781118901731.iecrm0011

Ahmed, S. (2015). *The Cultural Politics of Emotion*. New York, New York: Routledge. doi:10.4324/9780203700372

Bal, M. (1999). *The practice of cultural Analysis: Exposing Interdisciplinary INTERPRETATION*. Edited by Mieke Bal. Retrieved June 2, 2019, from http://www.sup.org/books/title/?id=851

Banet-Weiser, S., Miltner, K. M. (2015). Masculinitysofragile: Culture, structure, and networked misogyny. *Feminist Media Studies*, 16(1), 171-174. doi:10.1080/14680777.2016.1120490

Barker, C. (2004). Genre. In *The Sage Dictionary of Cultural Studies* (pp. 74-75). London: Sage.

Barker, C. (2004). Subjectivity. In *The Sage Dictionary of Cultural Studies* (pp. 194-195). London: Sage.

Bauman, Z. (2018). *Liquid modernity*. Cambridge: Polity Press.

Benedict, R., Goldenweiser, A. (1937). Patterns of culture. *American Sociological Review*, 2(5), 802. doi:10.2307/2083847

Berlant, L. (2011). *Cruel Optimism*. Duke University Press.

Blackmore, S. J. (2000). *The Meme Machine*. Oxford: Oxford University Press.

Blackmore. (2001). Evolution and memes: The human brain as a selective imitation device. *Cybernetics and Systems*, 32(1-2), 225-255. doi:10.1080/019697201300001867

Boellstorff, T. (2012). *Ethnography and virtual worlds: A handbook of method*. Princeton: Princeton University Press.

Borenstein. (2004). Survival of The Catchiest: Memes and Postmodern Russia. *The Slavic and East European Journal*, 48(3), 462-483. doi:www.jstor.org/stable/3220071

Brinkema, E. (2014). *The forms of the affects*. Durham: Duke Univ. Press.

Clough, P. T. (2008). The affective turn. *Theory, Culture Society*, 25(1), 1-22. doi:10.1177/0263276407085156

Coker, C. (2008). War, memes and memeplexes. *International Affairs*, 84(5), 903-914. doi:10.1111/j.1468-2346.2008.00745.x

Comaroff, J., Comaroff, J. (2003). Ethnography on an awkward scale. *Ethnography*, 4(2), 147-179. doi:10.1177/14661381030042001

Conte, R. (2001). Memes through (social) minds. *Darwinizing CultureThe Status of Memetics as a Science*, 83-119. doi:10.1093/acprof:oso/9780192632449.003.0005

Cvetkovich, A. (2012). *Depression: A public feeling*. Durham, North Carolina: Duke University Press.

Dağtaş. (2016). 'Down With Some Things!' The Politics of Humour and Humour as Politics in Turkey’s Gezi Protests. *Etnofoor*, 28(1), 11-34.

Dennett C. (1990). Memes and the exploitation of imagination. *The Journal of Aesthetics and Art Criticism*, 48(2), 127-135. doi:10.2307/430902

DeWalt, K. M., DeWalt, B. R. (2011). *Participant observation: A guide for fieldworkers*. Lanham, Maryland: Altamira Pres, Md.

Falzon, M. (2016). *Multi-Sited Ethnography: Theory, Praxis and Locality in Contemporary Research*. Routledge.

Gal, N., Shifman, L., Kampf, Z. (2016). “It gets better”: Internet memes and the construction of collective identity. *New Media Society*, 18(8), 1698-1714. doi:10.1177/1461444814568784

Geiger, R. S., Ribes, D. (2011). Trace Ethnography: Following Coordination through Documentary Practices. *2011 44th Hawaii International Conference on System Sciences*. doi:10.1109/hicss.2011.455

Godwin, M. (1994, January 10). Meme, Counter-Meme. *Wired*. Retrieved June 2, 2019, from www.wired.com/1994/10/godwin-if-2/

Highmore, B. (2008). Introduction: Questioning everyday life. In *The everyday life reader* (pp. 1-34). London: Routledge.

Hjorth, L., Cumiskey, K. M. (2018). Mobiles facing death: Affective witnessing and the intimate companionship of devices. *Cultural Studies Review*, 24(2), 166-180. doi:10.5130/csr.v24i2.6079

Hjorth, L., Pink, S. (2013). New visualities and the digital wayfarer: Reconceptualizing camera phone photography and locative media. *Mobile Media Communication*, 2(1), 40-57. doi:10.1177/2050157913505257

Hjorth, L., Horst, H. A., Galloway, A., Bell, G. (2017). *The Routledge Companion to Digital Ethnography*. New York: Routledge.

Holdcroft, D., Lewis, H. (2000). Memes, minds and evolution. *Philosophy*, 75(2), 161-182. doi:10.1017/s0031819100000231

Jameson, F. (2018). Postmodernism and consumer society. In V. B. Leitch (Author), *Norton Anthology of Theory and Criticism* (pp. 1758-1771). New York: Norton Company Limited, W. W.

Jenkins, H. (2009, February 11). If it doesn’t spread, it’s Dead (part ONE): MEDIA viruses and memes. Retrieved June 2, 2019, from http://henryjenkins.org/blog/2009/02/if_it_doesnt_spread_its_dead_p.html

Jorgensen, D. L. (2015). Participant observation. *Emerging Trends in the Social and Behavioral Sciences*, 1-15. doi:10.1002/9781118900772.etrds0247

Kolářová. (2017). The inarticulate Post-socialist Crip on the cruel optimism of NEOLIBERAL transformations in the Czech Republic. *Culture - Theory - Disability*. doi:10.14361/9783839425336-013

Lefebvre, H. (1992). *Critique of everyday life* (Vol. 1) (J. Moore, Trans.). London: Verso.

Leys. (2011). The turn to affect: A critique. *Critical Inquiry*, 37(3), 434-472. doi:10.1086/659353

Malinowski, B. (2015). *Argonauts of the western pacific*. Routledge.

Marwick, A. (2013). Memes. *Contexts*, 12(4), 12-13. doi:10.1177/1536504213511210

Murthy. (2008). Digital ethnography: An Examination of the Use of NEw Technologies for Social Research. *Sociology*, 42(5), 837-855. doi:10.1177/0038038508094565

Narayan, K. (1993). How native is a "Native" Anthropologist? *American Anthropologist*, 95(3), 671-686. doi:10.1525/aa.1993.95.3.02a00070

Ngai, S. (2007). *Ugly feelings*. Cambridge, Massachusettes: Harvard University Press.

Ngai, S., Berlant, L. (2017). Comedy has issues. *Critical Inquiry*, 43(2), 233-249. doi:10.1086/689666

Nissenbaum, A., Shifman, L. (2015). Internet memes as contested CULTURAL capital: The case of 4chan’s /b/ board. *New Media Society*, 19(4), 483-501. doi:10.1177/1461444815609313

Postill, J., Pink, S. (2012). Social media ETHNOGRAPHY: The DIGITAL researcher in a messy web. *Media International Australia*, 145(1), 123-134. doi:10.1177/1329878x1214500114

Puar, J. (2012). Precarity Talk: A Virtual Roundtable with Lauren Berlant, Judith Butler, Bojana Cvejić, Isabell Lorey, Jasbir Puar, and Ana Vujanović. *TDR/The Drama Review*, 56(4), 163-177. doi:10.1162/dram$_{a0}0221$

Rossolatos, G. (2014). The ice-bucket challenge: The legitimacy of the MEMETIC mode of CULTURAL reproduction is the message. *SSRN Electronic Journal*. doi:10.2139/ssrn.2505616

Saito, A. P. (2017). Moe and internet MEMES: The resistance and accommodation of Japanese popular culture in China. *Cultural Studies Review*, 23(1). doi:10.5130/csr.v23i1.5499

Sampson, T. D., Maddison, S., Ellis, D., Seigworth, G. J. (2018). *Affect and social media emotion, mediation, anxiety and contagion*. London: Rowman Littlefield International.

Schroeder, R. (2018). *Social theory after the internet: Media, technology and globalization*. London: UCL Press.

Shifman, L. (2014). *Memes in digital culture*. The MIT Press.

Shifman, L. (2014). The cultural logic of PHOTO-BASED MEME GENRES. *Journal of Visual Culture*, 13(3), 340-358. doi:10.1177/1470412914546577

Sterelny, K. (2006). Memes revisited. *The British Journal for the Philosophy of Science*, 57(1), 145-165. doi:10.1093/bjps/axi157

Taussig, M. T. (2016). *The Nervous System*. London: Routledge.

Zenner, E., Geeraerts, D. (2018). One does not simply process memes: Image macros as multimodal constructions. *Cultures and Traditions of Wordplay and Wordplay Research*, 167-194. doi:10.1515/9783110586374-008

Humanities

# Pieces of Resistance: Protest Signs as Objects of Dissent

Nina Klaff

*Supervisor*
Dr. Erinç Salor (AUC)
*Reader*
Dr. Marco de Waard (AUC)

**Abstract**

While much has been written about protest and its place in the new media age, little research makes use of visual methods. I argue the importance of the use of visual analytic processes to understand protest culture by investigating protest signs as politico-cultural objects. I argue that the protest signs made and shown by protesters themselves are artifacts of individual expression through which people perform both their political and personal identities. I first present a firm theoretical grounding of protest marches. The research methodology includes attendance of protest events, as well as cultural and formal analyses of their artifacts. Two case studies are analysed. First, the People's Vote March, demanding a second vote on the U.K.'s Brexit Referendum, provides a basis for investigating how protest culture has evolved with social media by analysing the rhetoric and aesthetics employed in the signs to draw attention to their cause. Second, the Global Women's Marches (2017-present), an international movement of feminist marches, demonstrates how movements and their signs travel transnationally through social media. I then examine political statements made through visual media in online identities to argue that the space of appearance itself is digitally networked and that our conception of protest signs can be applied to other forms of aesthetic expression, which include those created in the digital sphere.

Keywords and phrases: *materiality, embodiment, performativity, aesthetics, assembly*

### *A note on form*

*Parts of this paper are punctuated with short texts in black boxes. These are words, slogans, sentences I have seen on protest signs. These have been placed to interrupt the reading process, their messages becoming part of the spectacle of this paper, to emulate the prominence of signs in the constellation of a demonstration. Inspired by Michael Taussig's "I'm So Angry I Made A Sign," in which he includes some texts from signs held by protesters at the Occupy Wall Street occupation of Zuccotti Park, New York (2011) in these same black boxes. This strips them of their aesthetic, embodied, and material dimension, and lets the words speak for themselves. Consider what has been lost, and what has been gained, by doing so. It is hard to give credit to each of their authors, so instead I attribute credit to the movement within which they were shown. These signs are here to be read alongside the words of other scholars, researchers, and people of note. As Taussig said, "I don't think you will confuse them, but it's better that you do" (56).*

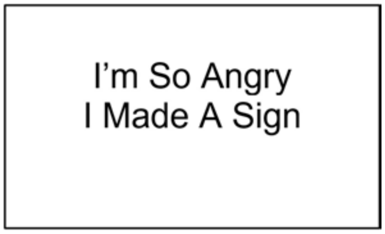


*Figure 1: Klaff, Nina. Illustration of a sign at Occupy Wall Street as reported in Taussig's "I'm So Angry I Made A Sign," 2019.*

## Introduction

On the 20th of October 2018, 700,000 people marched the streets of London demanding a second vote on the Brexit Referendum that entailed the U.K.'s leaving of the European Union (*BBC News*). I was one of them. While the gathering of so many bodies in the public space was, as it was meant to be, a beguiling carnivalesque spectacle, the most striking visual components of this march were the signs made and carried by the protesters themselves. The prevalence of protest signs is not particular to this event. It is traditional in protest culture to carry signs that demonstrate one's participation within a protest event, which express the views of the individuals who carry them as well as the character of the movement itself. Protest signs signal not only participation within a particular movement, but are also signals of dissent themselves.

In Charlie Chaplin's *Modern Times* (1936), Chaplin is seen walking down the street among a montage of modernity: factory machinery, cars, and crowded streets. A truck drives past him, a flag perched precariously on the back. As the truck drives off, the flag falls. Chaplin runs into the road and picks it up. As he waves it in the air to alert the driver, a crowd appears behind him. Chaplin is oblivious. The crowd is holding signs. As he continues to chase the car, waving the flag, viewers are made to understand that he has inadvertently found himself at the head of a protest march as the police exclaim, "so, you're the leader!" (*Mod-*

*ern Times* 1936). On his own, Chaplin was a mere passer-by. The flag, left on the back of the truck, or lying on the street, held little meaning. However, as soon as Chaplin picked it up, its symbolism was catalysed, even without it carrying any textual or other symbolic cues. The flag made a protester of Chaplin, otherwise an uninvolved bystander; and Chaplin, by inadvertently joining an assembly, made a protest sign out of an otherwise insignificant flag.

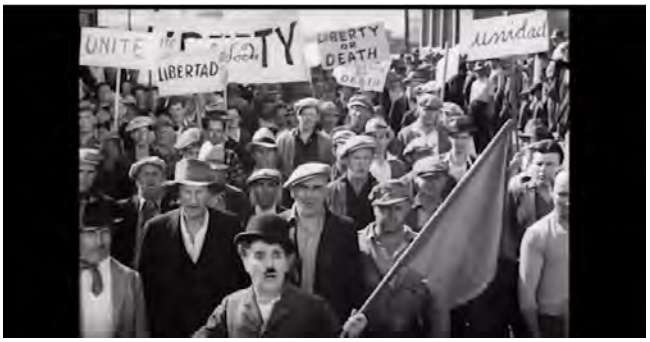


*Figure 2: Still from Modern Times, 1936, retrieved via YouTube. "Charlie Chaplin. Modern Times. Protest Sign." YouTube, 00:54, uploaded by user GeslominaAM, 10 Mar. 2014. Accessed 30/05/2019.*

The embodied fusion of Chaplin and this object gave political significance to them both. The mere term *protest sign* demonstrates that these objects function as universal signals of protest; these are important objects of political and human history in popular culture.

In *Twitter and Tear Gas*, Tufekçi demonstrates that "collective actions, social movements, and revolutions are woven into the fabric of human history. They have been studied at great length and for good reason: they change history" (xix). However, as Philipps argues, their scholarship "rarely employs visual analytic procedures" (3). I argue that this is a drastic oversight. The visual objects deployed by protest movements give great insight into their nature and culture, and are themselves objects of political significance. As Khatib states, "the image is at the heart of political struggle," which, she argues, is "an inherently visually productive process" itself, because it "is a struggle over presence, over visibility" (qtd. in Werbner 2). Thus, visual devices, such as protest signs, are important political objects.

As Sartwell argues, "not all art is political, but all politics is aesthetic" (qtd. in Werbner 1). Protests are political manifestations, and are thus, in this understanding, also aesthetic. I understand aesthetic under Rieff's definition: "that activity in which, by means of external signs, human beings communicate feelings to each other" (479). I thus pay due attention to protest signs as external devices used by 'the people' to communicate their politics. The signs act as expressions of the thoughts and beliefs of the carriers, who use them to signify what they want to say. Protesters make them by hand, by writing, painting, drawing, or sometimes printing a statement onto paper or cardboard. They decorate them with images and other symbols for the purpose of holding them up and carrying them at protest events. Sign-holding is an embodied, performative, material, and aesthetic action of political expression. I argue that these signs, imbued with political meaning, are not only important objects for visual and cultural analyses, but are also emblematic of protest culture and that their production and display in protest activities is a revolutionary practice itself. I first establish a firm theoretical framework, drawing from protest discourse in the fields of visual culture, anthropology, sociology and materiality, to explore how contemporary protest movements express their aims (e.g. Butler *Notes on Assembly*; Reiss textitThe Street as Stage). Drawing from Tufekçi's analysis of 21st century social movements, I argue that protest signs allow a movement to set their own narrative, and thus epitomise their "*narrative capacity*" (192). I build on Mitchell's seminal question, "what do pictures want?" to posit the consideration of *what protest signs want* (71). I further previous protest scholarship to establish what is meant by the space of appearance and how movements use protest marches to express their aims (Arendt *The Human Condition*).

I then explore the innovation in contemporary protest signs by analysing their role within two movements.[1] Firstly, the People's Vote marches in London in 2018 and 2019, which demanded a second vote on the Brexit referendum, were limited to nation-specific politics. I analyse the movement's signs in terms of materiality and politicised humour, which entails a consideration of

[1] The two case studies in question (the People's Vote March and the Women's March) are non-violent, peaceful, and democratic dissenting events. Of course, other forms of protests feature signs, but for the sake of the clarity and brevity of this paper, I focus mainly on these protest marches.

social media aesthetics. To facilitate an analysis of how social media might be utilised by contemporary movements, I consider the signs at the Global Women's Marches (2017-present): a transnational movement that exemplifies how social media has changed the organisation, execution, and the very nature of contemporary protest movements, demonstrating that our conception of protest culture must now include the digital sphere. Such an analysis serves to highlight the ubiquity of protest signs in contemporary protest culture, and encourage an understanding of their importance as objects of cultural interest and of socio-political import.

## I. Placing Signs in Protest

In *Twitter and Tear Gas*, Tufekçi defines a social movement as "a claim made to a public that a wrong should be righted or a change should be made" (8). I thus propose a consideration of the ways social movements use protest signs to demand such change. Tufekçi notes how important it is to recognise the risks and dangers associated with protest in certain contexts (Tufekçi 87). In democratic societies such as Britain, the Netherlands, and the United States, demonstrations in the public sphere remain legal, and entail less risk of danger. However, both the People's Vote Marches and the Women's Marches in part arose out of the fear of such liberty being curtailed under new governments and legislations. Thus, even in peaceful and non-violent protests, the stakes remain high. Moreover, as Tufekçi argues, "protests have always had a strong expressive side, appealing to people's sense of agency" (8). Furthering this, I argue that protest signs are emblematic of the expressive qualities of a movement, as they facilitate the creative expression of individual agency. Tufekçi demonstrates that markers of a social movement's success and strength "are more complex than indicators like headcounts or number of protests" and outlines "three crucial capabilities of social movements from the point of view of power: narrative capacity, disruptive capacity, and electoral and/or institutional capacity" (191-2). Under this definition, "narrative capacity refers to the ability of the movement to frame its story on its own terms, to spread its worldview;" "disruptive capacity" defines whether the movement can "disrupt the status quo;" and "electoral or institutional capacity," refers to a movement's ability to incite electoral or institutional change (Tufekçi 192). Under this definition, "disruptive capacity" outlines whether the movement can "disrupt the status quo;" "electoral or institutional capacity," refers to a movement's ability to incite electoral or institutional change; and "narrative capacity refers to the ability of the movement to frame its story on its own terms, to spread its worldview" (Tufekçi 192). I argue that while the movements in point might not have overtly disrupted the status quo or incited tangible electoral or institutional change, they have been effective in setting their own narrative. The protest sign was one of the key devices deployed to do so.

I argue that these objects contribute invaluably to the "*narrative capacity*" of social movements. They are objects of what Jenkins might call "participatory culture," whereby people actually contribute and participate in the production of communication (qtd. in Gerbaudo 22). They allow each sign-maker to set their own personal narrative, which in turn shapes the narrative of the movement itself. Tufekçi argues that speaking up "in a challenging way in public" tends to be the reserve of the privileged (100). [2] Protest signs enable people who might not otherwise feel comfortable to speak up, to carry their own voices, loud above their heads, and parade them through the public sphere, demonstrating their own convictions and potentially persuading others of their beliefs. As Garlough argues, "rhetoric is the art of persuasive speaking or writing, [but] it can be applied to other forms of aesthetic performances that aim to persuade" (272). Protest signs, as aesthetic objects that express political views, can be considered for their performative rhetoric of political persuasion.

### Texts and their rituals

It is first important to consider signs in the context in which they are displayed. Exploring the materiality of text-based rituals in Judaism, Stolow argues that "texts are governed by various protocols that make them perceptible; engagement with them is not only intellectual but also embodied" (316). He outlines the rituals surrounding the creation and production of the Torah scroll, such as purification rituals that scribes must observe (319).

[2] As she also notes, in her own experience, it also tends to be the province of men.

These contribute to the religious significance of the holy text in its material form. I use this illustration to demonstrate that there are rituals surrounding textual objects. Protest signs become perceptible when they are displayed at protest events. Thus, any analysis of signs requires a consideration of the context in which they are displayed. As Tufekçi argues, protest ceremonies have their own rituals, which she terms a movement's "repertoires of contention" (89). The protest sign is a text that has its own rituals of production and use. It is customary to make signs, and carry them at marches, holding them up high, marching them through the streets. They have become as much a part of the march as the people who gather together in participation, and integral textual objects in the ritual of protest marches.

## Signs in Context: Protest Marches

In *Dissensus*, Rancie`re argues that "politics begins when those who were destined to remain in the domestic and invisible territory of work and reproduction, and prevented from doing 'anything else,' take the time that they 'have not' in order to affirm that they belong to a common world" (139). This is what happens at protest marches: a group of people takes the time to come together, united on the common space of the streets, to affirm their unity of thought on a certain issue - which is most often political.

Protest events offer, as Reiss argues, "the opportunity for social or political movements to showcase themselves to the public in various ways and solicit support" (3). The idea of a showcase is of particular interest as the language of protest discourse is permeated with the lexicon of performance. For the sake of this research, I focus on two protest marches that were peaceful, non-violent performances of dissent in the public sphere. Marches are a "relatively new phenomenon:" until the 20th century, other forms of public representation were favoured (Reiss 2). Protest marches are "understood as organised and choreographed processions of groups in the public sphere with the aim of making a statement" (Reiss 2). The street is an ideal stage for protest events in its symbolic meaning as a fundamentally public space and its material affordance as a facilitator for the movement of bodies from one place to another. As Reiss argues, "the street is a stage on which feelings and convictions can be expressed, issues addressed, demands presented, and support solicited" (2). A people is ruled by those who rule the streets upon which they walk. By gathering on the streets, they reclaim that public sphere as their own, thus symbolically reclaiming agency over their governance.[3] It is a demand for their presence in society to be acknowledged, for their demands as political agents to be heard, and for their shared desires as individuals within a larger group to be recognised.[4]

Thus, if the street is to be thought of as a stage, we can begin to understand protest activities as performances on that stage and ascertain how the devices used as props to that performance might function.

## The Embodied Assembly

What marks the difference between a protest placard and a banner or poster that is simply hung out of a window, displayed on the street, or plastered on a wall, is that signs are held by protesters in the context of demonstrations. Understanding their significance thus requires careful consideration of embodiment in protest rituals and an understanding that the gathering of bodies in the public sphere is itself a political act. It is the performance of "the right to appear, a bodily demand for a more livable set of lives" (Butler 25). Indeed, one merely has to cast an eye over the history of political activism to conclude the importance of the human body in socio-political contention (Kraidy 2018). By participating in a march, as Reiss posits, "the marchers embody not only the issue they represent, but also the movement behind it" (3). This embodied co-presence of people in the public sphere is a bodily show of popular dissent in the lexicon of politics. Butler reminds us that an assembly speaks volumes: "the coming together of a crowd has, as John Inazy contends, 'an expressive

[3] Mitchell, in his analysis of the Occupy Wall Street movement, turns our attention to the empty, negative space against which assemblies appear and argues that the demand of "occupatio," of embodied presence in the public sphere, is made in the full knowledge that public space is, in fact, pre-occupied by the state and the police" (10).

[4] I here refer only to democratic societies, such as the contemporary UK and United States, where protest marches remain legal. Indeed, the state and the police is often given fair warning of such events, in order for the proper preparations to be put in place. There is an interesting paradox in a government actually allowing and protecting citizens even as they voice dissent against them.

function’ prior to any particular claim or utterance it may make” (161). There is a powerful statement being made each time people gather, as they do, to express themselves, which is heightened when it is placed in the public sphere. As Butler argues, “*the assembly is already speaking before it utters any words*” (156). An assembly itself is a demonstration of solidarity and a way of communicating a people’s message. It is a show of collective unity between individual people, united in a singular aim who are willing to put their bodies on the line to make a political statement.

I argue that holding a protest sign not only emphasises the claim to the right to appear, but also substantiates that claim as well as furthering what is being expressed in an individual’s participation in an embodied protest. Arendt marks an important difference between individual strength and collective power: “while strength is the natural quality of an individual seen in isolation, power springs up between men when they act together and vanishes the moment they disperse” (200). Power, according to Arendt, necessitates a multiplicity of people: the crowd invokes the power of the people in the strength of numbers and through their union in one aim, and it is only when “men” - or indeed just people - “are together in the manner of speech and action,” that “the space of appearance comes into being” (199).

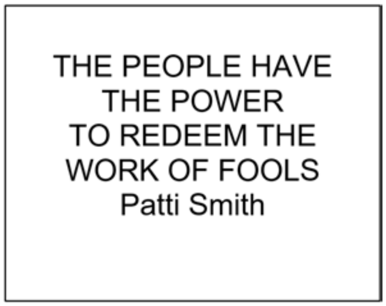


*Figure 3: Klaff, Nina. Illustration of Patti Smith sign seen at the People’s Vote March, March 2019.*

Indeed, the more people are present, the louder their claims are made. Thus, while one individual, or sign, might have strength, it is through their display in communion in numbers that the power of the voices they carry is truly enacted. Disregarding for a moment any other devices might be used to express political claims and demands, the mere gathering of bodies in the public sphere creates and sustains the space of appearance, which Arendt argues is “wherever people gather together, it is potentially there, but only potentially, not necessarily and not forever” (199). Its embodied nature entails a materiality that is of great import. The space of appearance is thus not just public space, but is also created and sustained by the assembly of bodies. The “space of appearance,” she contends, “unlike the spaces which are the work of our hands,” “disappears with the dispersal” of those that created it (Arendt 199). The collective communion of bodies is thus a significant site of political dissent.

Durkheim coined the term “collective effervescence” in 1912 to explain the ebullient energy that emerges in a gathering of people which is the feeling of being “swept through one’s actions by a larger power when one is ‘within a crowd moved by a common passion’” (Durkheim quoted in Tufekçi 89). Durkheim was speaking then of religious life; however, as Tufekçi notes, he himself “showed that rituals also extend to secular activities” (89). In Durkheim’s own words, “the very fact of assembling is an exceptionally powerful stimulant” (Durkheim qtd. in Gerbaudo 39). “*Collective effervescence*” can be felt whenever there is a crowd of people gathered together, united in one singular aim, belief, or ritual: as Tufekçi posits, “that transcendent feeling of being part of something larger than oneself applies also to protests, however secular their aims may be” (89). The protest march is an embodied tradition, a ritual of dissent, in which an assembly of individuals gathers together to demonstrate their unity as a group, united in a common aim, and “this relationship between belonging and individual expression is a key component of protest participation” (Tufekci 89). It is within this coming together, by gathering in the public sphere, in the creation of the space of appearance, that individual freedom, and indeed, dissent, is expressed, performed, and embodied. I argue that protest signs heighten “collective effervescence” as they signifies participation, which entails a feeling of belonging, and that these are, most importantly, objects of individual expression.

## Protest Marches as Carnivalesque Spectacles

Marches are performative spectacles; and as Werbner states, “voting lacks the spectacular aspects of packed bodies in the square” (10). These “revolutionary moments,” Vaneigem argues, “are carnivals in which the individual life celebrates its unification with a regenerated society" (qtd. in Tancons 297). The analogy of the carnival, the popular festival, is also often employed in protest discourse: these "festivals of resistance" are empowering occasions that, through their carnivalesque qualities, might fortify one's personal identity while simultaneously consolidating one's feelings of belonging (Graeber 23). Werbner et al. remind us that "as Bakhtin showed for the medieval carnival, carnivals are contained moments of 'rebellion:'" against the rigidity of the quotidien, a popular celebration of *us* against *them* (12). Tancons, however, warns of the danger of the carnivalesque analogy undermining protest activities, and maintains that the analogy of the carnivalesque is not meant to undermine the gravitas of a movement's claim: "carnival is serious business" (Tancons 291). Carnival is a revolutionary performance practice that turns the status quo on its head and the performance of protest activities are not to be thought of as in any way frivolous. In Tate's words, people "don't do demos or entertain police assault for abstract carnivalesque goals" (quoted in Tancons 302). Simply by gathering in the public sphere, individuals make their potentially overlooked presence in society overtly visible. Protest marches are feasts for the senses, and placards add to the visual spectacle. Through them, protesters can demonstrate their otherwise private individual political thoughts and emotions. Rancière defines politics as when people "make the invisible visible, and make what was deemed to be the noise of suffering bodies heard as a discourse concerning the 'common' of the community" (139). I argue that protest marches serve to make the potentially invisible aims and desires of its participants visible, and that what can differentiate between this noise and the discourse concerning the 'common' is an object as simple as a protest sign.

## Materiality

The protest sign is that which links the symbolic assertions, claims and demands made by the attendance of a march with the claim of rights and freedom of expression made simply through the creation of the space of appearance by the gathering of bodies. Within these signs, the strength of each individual protester can be read, and when gathered together they represent the power and the character of the movement itself. Moreover, the emerging lens of materiality offers interesting perspectives on cultural objects such as protest signs: as Trentmann proposes, "after the turn to discourse and signs in the late twentieth century, there is a new fascination with the material stuff of life" (283). Under this light, we may observe the materiality of protest signs themselves: these are often stuck on wooden or plastic sticks and bars to hold above the body. Activists might gather before a march to create their signs together, in solidarity with each other, and the signs move with the protesters through the crowd, as much a part of the constellation of the event as the bodies gathered together of which it is contingent. Taussig, observing the Occupy Wall Street protesters in Zuccotti Park, sees "the sign as an extension of the human Figure," and the protester/sign fusion as "centaur-like" (75). An understanding of the symbolism they carry thus requires a consideration of their nature as material objects in their formal analyses, such as in Chapters One and Two.

## Mediation

In contradiction to Arendt, I propose that the spectacle does not end when the crowd disperses. For De Kloet, walking through the streets of Hong Kong during the Umbrella Revolution, "the movement felt like a theater show, a spectacle of resistance, built not only to be experienced, but above all to be mediated" (159). The aim of protest events is to interrupt the flow of daily life by creating a spectacle that will encourage the communication of the reasons for the performance of dissent, in the hope that this will subvert the status quo. Werbner et al. remind us of the work of Juris, who "proposes that these 'image events' communicate to wider audiences by 'hijacking' the global media, while at the same time creating affective solidarity through performance" (11). This consideration of the attention drawn by these protest events used as image events and the solidarity created through performative devices leads to a consideration of the aesthetic inventions that people use as props to their performance of protest. As Butler recognises, "every activist needs to negotiate how much exposure, and in what way, is necessary to achieve his or her political goals" (55). The ethical concerns of media framing - who gets to be seen and in what way - is a major concern, but as Serafini argues, social media "have enabled activists to record and then share creative spectacles as never before” (321). Moreover, activists can control how an event itself appears by peppering the spectacle with signs that clearly say what they want to say, and in their own voices. Indeed, Taussig argues that in a protest event, “the sign holder is posing for photographers, or, rather, the sign is being made to pose for the camera” (Taussig 75). The protest sign is no longer, nor perhaps has it ever been, confined to the embodied spectacle, but is an object that demands to be remembered, mediated, and shared.

### What Do Pictures Want?

In his seminal 1996 essay, Mitchell turns visual culture traditions on their head by displacing the notions of desire that are usually read into images and returning agency to the images themselves. He asks: “What Do Pictures Really Want?” and concludes, through tracing visual culture “away from meaning and power to the question of desire”, that “what pictures want in the last instance, then, is simply to be asked what they want, with the understanding that the answer may well be, nothing at all” (Mitchell 81-82). Protest signs as political objects make political demands; they don’t ever want *nothing*. They demand that the voices of the people be listened to, and this depends on the individual and the movement within which they are shown. Stolow argues that texts, as they are “materially embodied,” in books or digital media, or even in protest signs, “possess a sort of agency, demanding from their human users the acquisition of certain competencies and skills, and enabling various modes of performance and action” (318). I thus use Mitchell’s question in specific relation to placards, and ask *what do protest signs really want?* I argue that the desires of protest signs, as images and texts at once, can only be read in the context of the movement within which they are used. We must first ascertain what the movement is asking for.

Before we can adequately answer what protest signs want, we first have to address what the gathered assembly itself wants, which common passion unites them. The gathering of people in the public sphere makes demands in itself, but the explicit aims of that assembly are particular to each movement. I posit a consideration of the elements intrinsic to protest marches as rituals of expressing political dissent, such as protest signs, as an entry point into the understanding of the importance of protest signs as cultural objects. I thus propose a consideration of the People’s Vote March and the Women’s March to demonstrate how the aims and character of a movement can be read in the protest signs they display.

## II. The People’s Vote March

*Figure 4: Illustration of a sign seen at the People’s Vote March, 23rd March 2019.*

### Context

In June 2016, the United Kingdom held an advisory referendum on its membership of the European Union. The British public voted to leave the EU, “forcing the resignation of Prime Minister David Cameron and dealing the biggest blow since World War Two to the European project of forging greater unity” (Faulconbridge). The Leave campaign, popularised by then UKIP leader - now leader of the Brexit Party - Nigel Farage, won with 51.9% of the votes, to Remain’s 48.1% (BBC News). The marginal outcome of the referendum has been the subject of long-running contention, with the Leave campaign then under legal scrutiny by the British High Court following allegations of “significant overspending, data breaches and possibly Russian involvement in the referendum” (Bowcott). These “corrupt and illegal practices” have been ruled by the Court to “undermine the validity of the decision to leave the EU” in a referendum that remains advisory rather than regulatory (Bowcott). Nonetheless, Theresa May’s government continued to persist in its attempts to negotiate a deal. The People’s Vote organisation, led by James McGrory, has been campaigning for a second referendum that they believe would yield a different outcome. The movement gathered an estimated 700,000 people on the 20th of October 2018, and over a million again on the 23rd of March 2019 who marched the streets of London in the demand of a People’s Vote. An analysis of this movement serves to give insight into some of the aesthetics, rhetoric, and power at play in contemporary protest signs.[5]

## Materiality: The Streets, The Signs

A city embodies the history and culture of its citizens and as Kapferer notes, "as a built formation, the city materializes, houses, and gives substance to its social and ideological structure" (69). I here consider protest activities through the lens of materiality to demonstrate how the concrete structures of cities and their histories and geographies are used by political movements as platforms upon which to express dissent through other material devices. As Tufekçi argues, cities "alter how we interact by gathering people in large numbers and creating places for interaction outside of private spaces" (5). The People's Vote Marches took place in London: the importance of its location lies not only in this being the capital of the United Kingdom: by considering the geography of the city and its landmarks, we may regard the route chosen as a symbolic reclaiming of the social and ideological structure of the capital. Kapferer conducts an in-depth analysis of the socio-political history of London's monuments, which she posits "as crucibles for the expression, symbolization, formation, and re-formation of the social orders of the city and the state" (Kapferer 67). An examination of the geographical setting of the People's Vote march and the landmarks they encountered and occupied will serve to illustrate that the message conveyed by assemblies is heightened against the backdrop of concrete symbols.

Both the march on the 20th October 2018 and that on the 23rd March 2019 started in the same place - and at the same time of day. Protesters first gathered on Park Lane, a road synonymous with affluence, that borders Hyde Park - London's iconic Royal Park - and its vast spread of greenery adjoins London's wealthiest and most influential areas.

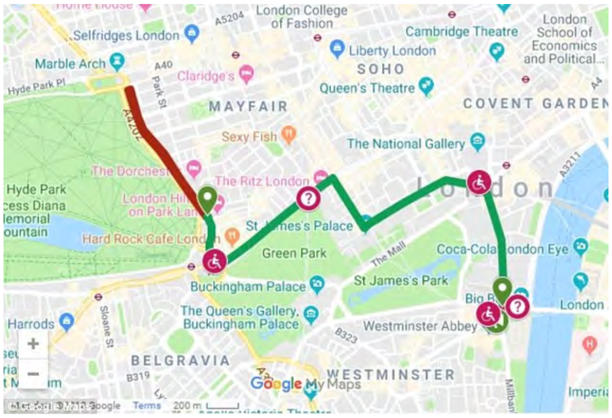


*Figure 5: Ben Rabinovich. People's Vote March route, Google Maps, 19 Oct. 2018. Accessed 14/03/2019.*

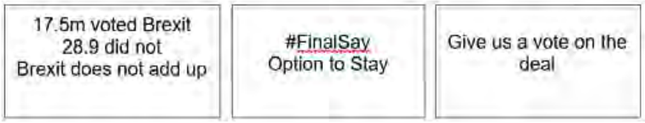
To access Park Lane by tube most conveniently, one must alight at Marble Arch station, a stone's throw from the famous Speaker's Corner, which Kapferer describes as "traditionally a venue for the expression of popular debate and the airing of contemporary contention" (77). Protesters then moved to Picadilly, a landmark London road dotted with high-end shops and hotels such as the famous The Ritz, and then went down St James's Street, past St James' Palace, one of the oldest royal buildings in the city, to Trafalgar Square. Kapferer reminds us that Mace calls the latter "The Heart of the Empire," and herself describes its aura as "half-tourism, half-ideology" (69; 75). Of the square, Kapferer says "first and foremost, it has long been the great gathering ground of the British people in times of triumph, jubilation, tribulation, and rage" (69). Its monuments, namely Nelson's Column, "all unequivocally bespeak the power of an imperial past" (Kapferer 75). Its central location, proximity to parliament, and dotting with monuments to historical Figures are significant in that each of these factors contribute to it being chosen as one of the most common backdrops for voicing dissent. Thus, it is in the concreteness of monuments, landmarks, and areas of a city, and their durability against the fragile mortality of the human body, that protests enact their symbolic value. This lens of materiality also facilitates a consideration of the materiality of protest signs.

[5] This analysis has been conducted in full recognition of the fact that every demonstration has many, many signs - especially one as large as the People's Vote March - and that no one person could ever witness, let alone analyse, every one of them. I have thus selected those I observed personally and deemed of particular interest for the purpose of this paper. However, so long as these objects are in production, so long will their analysis require revisiting and revising. I thus encourage further research into other signs carried there and at other marches.

## Aesthetic Regime

Protest signs are, first and foremost, material things. Fashioned out of found materials, these objects of artivism are made by the people, and this is reflected in their characteristic DIY aesthetic. Rancie`re argues that under the aesthetic regime of art, “art is art to the extent that it is something else than art” (118). Under this definition, I thus consider protest signs as forms of art in that that they also serve a political function. Indeed, Rancie`re asks us to reconsider our understanding of the “political import of artistic practices” (135). I argue that the creation of a protest sign is both an artistic practice and a political act of protest in itself. It is through the production of the objects that activists are able to enact their political autonomy and express their own individual political thoughts. Carrying them within the context of a protest march, they participate in the production of the spectacle and contribute to the character of the movement as a whole.

Some protest signs are straightforward, focusing simply on expressing the desires of the movement in a clear and focused manner, such as those in Figures 6-8.



17.5m voted Brexit
28.9 did not
Brexit does not add up

#FinalSay
Option to Stay

Give us a vote on the deal

*Figure 6;7;8: Illustrations of clear signs seen at People’s Vote March, March 2019.*

However, it has also become customary to put effort into producing high-quality signs that emphasise the aims of a movement and express an individual’s relationship with that movement in ever more innovative and engaging ways. There is a growing culture of producing effective protest signs that will capture the attention of the media and onlookers. Multiple media platforms published selections of “The Best Signs” shown at the People’s Vote March; from The Guardian’s “‘Fromage not Farage:’ the best signs and sights on the People’s Vote march,” to The Slate’s “People’s Vote March: The Best Signs From the Huge Anti-Brexit March in London,” the signs themselves can make headlines.

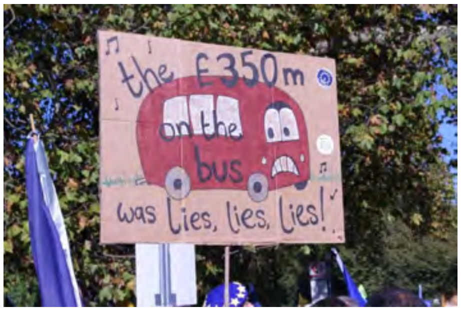


*Figure 9: Klaff, Nina. “The £350m on the bus was lies, lies, lies” sign, unpublished photograph, People’s Vote March, 20 Oct. 2018.*

In Fig.9, the sign is made of cardboard, attached to a stick of bamboo on which an anthropomorphised bus looks worried. The words “the £350m on the bus was lies, lies, lies!” are handwritten in felt-tip in different sizes, and musical notes such as quavers and beams decorate the background. This refers to the fact that in the months leading up to the 2016 Referendum, the Vote Leave campaign deployed a bus that traveled around the country plastered with the promise that £350 million paid by the United Kingdom to remain members of the European Union would be reallocated to saving the British National Health Service on the event of U.K. leaving the European Union (Getty Images).

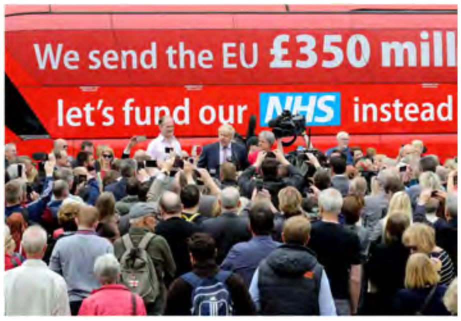


*Figure 10: Getty Images. Boris Johnson in front of the £350 million promise bus. The Independent, 8 Sept. 2018. Accessed 29/05/2019.*

Garlough argues that “playfulness and creativity, drawing upon news events, historical materials, folklore and popular culture” is a practice in many aesthetic productions of protest movements (270). On the placard in Fig.20, the musical notes, as visual symbols, imply a melody, and the bus, as well as the text written on it, is evocative of British nursery rhyme “The Wheels on the Bus,” in which the wheels go “round and round,” a line that is repeated to a sing-song melody. Thus, the bus subverts a song that many will remember from their childhood, combines it with the anger at a news event, and the subsequent effect is also humorous to an extent.

Fig.11 is a rather innovative sign. It is not made out of paper or cardboard but out of cloth, attached to two sticks, its message patchworked out of felt fabric and marker pen.

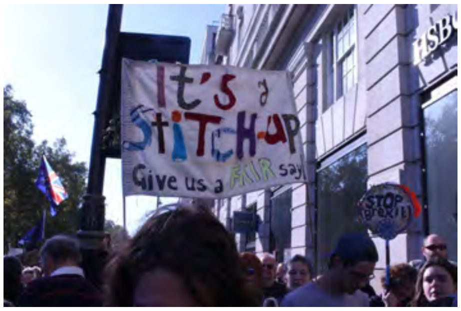
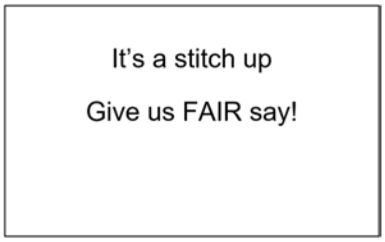


*Figure 11: Nina Klaff. It's A Stitch-Up Give us a FAIR say! Unpublished photograph, London, People's Vote March, March 2019.*

The sign has two sticks, one on either side, which means it requires being held up by two people to be visible. It is held way above the heads of the protesters. In this image, the sign is visible against the backdrop of the street, with a branch of the prominent British bank HSBC discernible in the background. Taussig proposes that "it is the handmadeness of the signs, their artisanal crudity, art before the age of mechanical and digital reproduction" that make them so impactful (75). Here, the "artisanal crudity" lies in its handmade quality. The sign is hand-stitched and handwritten; it declares the Brexit referendum to have been a "stitch-up," meaning that it was deceptive. Philipps argues that "visual methods can add insight to the analysis of text-based protest messages" (8). The sign in Fig. 8 deploys a self-referential visual and verbal pun uniting the medium and the message. However, this visual and material pun is lost in Fig.12, when the text is alienated from its original manifestation.



It's a stitch up

Give us FAIR say!

*Figure 12: Klaff, Nina. Illustration as transcript of "It's A Stitch Up" sign, People's Vote March, 2019.*

The "Stitch Up" sign is of arguably high quality, evidently well thought-out and executed. Protest signs are just one of the "countless examples of personal ingenuity" displayed in protest activities that aim to express opinion and solicit support for a movement (Tancons 296). Putting effort into the creation of a sign can indeed be a way of demonstrating one's commitment to a cause through creative aesthetic practice.

However, there is also a culture of resistance against this very phenomenon. Protesters have taken to making self-deprecating jokes about the limits of their signs (see Fig.13-14).

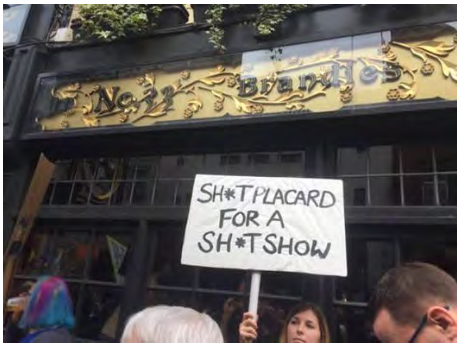


*Figure 13: Klaff, Nina. "Sh*t placard for a sh*t show." Unpublished photograph, People's Vote March, London, March 2019.*

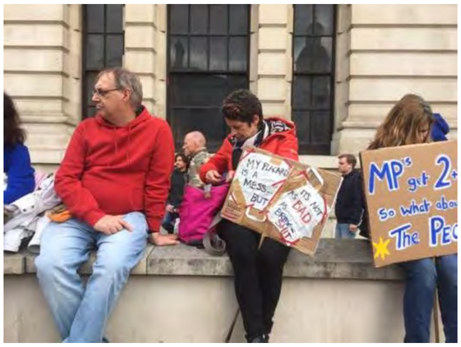
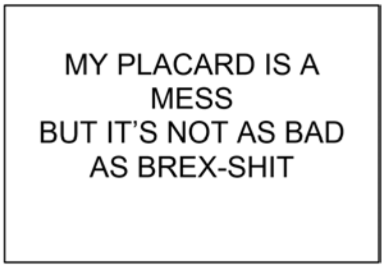


*Figure 14: Klaff, Nina. "My placard is a mess but it's not as bad as Brex-shit." Unpublished photograph, People's Vote March, London, March 2019.*



MY PLACARD IS A
MESS
BUT IT'S NOT AS BAD
AS BREX-SHIT

*Figure 15: Klaff, Nina. Illustration transcript of "Placard is a Mess" sign.*

However, even as the protesters use profanities to state they consider their placards artistically limited, they still objectively act as effective conveyors of dissent and signs of political contestation. What

Fig.13-14 might lack in artistic quality, they make up for in an important aesthetic quality: humour.

## Politicised Humour

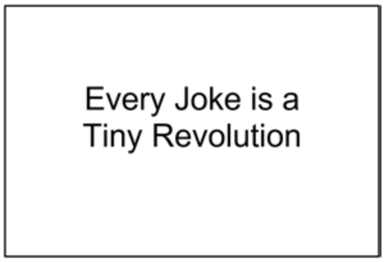


*Figure 16: Klaff, Nina. Illustration of "Every Joke is a Tiny Revolution," George Orwell (qtd. in Bruce 1).*

Contemporary protest signs often employ humour: through the use of quick quips, written and visual puns, they can be fun and playful objects of dissent. Holm argues that humour is "an unavoidable aspect of how we approach and understand the world as a site of meaning, politics, and life itself" (8). The humour in signs in no way trivialises protest activities, however. As Holm proposes that we think of humour itself as politics (43), the humour in protest signs functions politically. Holm establishes the notion of a "political aesthetics" of humour, which he argues demonstrates "the idea that the aesthetic aspect of a text - its form, style, palette, rhythm, narrative, structure and form - can do political work, by which [he means] it can intercede in the negotiation, contestation and distribution of power" (12). However, what we find funny is, of course, limited to particular contexts. Thus, Holm differentiates between humour, "an aesthetic quality operative at cultural level," and "funniness," which is "a particular subjective reaction to those texts" (19). Thus, I shall not address whether the protest signs are funny, but maintain that they often do contain humour, which can operate as a form of protest itself. Holm defines "humour that directly addresses the content of the political sphere," including but not limited to governments, is "politicised humour" (61). Thus, the humour in protest signs is politicised as they are objects that belong to the public sphere.

In a *New York Times* article about the use of jokes in the 2011 Egyptian revolution, Slackman states that "the organizers used humor as part of their communications strategy, to motivate people and bring out the crowds." Holm interprets Slackman's article as a demonstration of "the role of humour as a revolutionary tool" and argues that humour is thought "to exist as an entirely liberatory force in the aid of 'the people'" (Holm 41; 38).

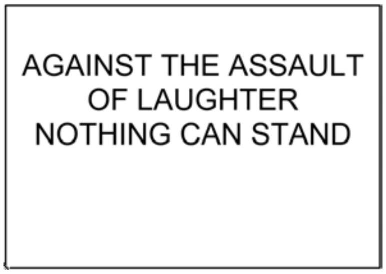


*Figure 17: Klaff, Nina. Illustration: Mark Twain, "Against the Assault of Laughter Nothing Can Stand" (qtd. in Holm 48).*

Just as the production and carrying of protest signs is an expression of independent thought in the constellation of group political dissent, the use of humour within those signs is also a revolutionary act. Holm reminds us that for Mindess, "humour operates as a key means by which a liberal subject might recognise and realise her own absolute autonomy" (43). In peaceful nonviolent demonstrations, in which political animosity is expressed through the gathered assembly and marching through the streets in a performative spectacle of protest, humour is a means of expressing individual autonomy. Holm argues that when humour is used in this way it "is humour as a force of freedom: not as an addendum, but as the very heart of a liberally focussed culture and politics of well-mannered and reasonable dissent" (42). I argue that protest signs contain humour in order to do positive political work, to reclaim the narrative and to subvert the status quo through subversion of authority in a reasonably well-mannered fashion.

I again stress that humour is serious political work. The aptly named Montaignian essay "Comme nous pleurons et rions d'une même chose" ("As we cry and laugh at one same thing," translation my own) demonstrates how outrage and despair can inspire comedic activities. In "The Anatomy of Schadenfreude," Carroll Simon works closely with Montaigne's theories of comedy. He works to demonstrate how laughter and crying, rather than being at opposite ends of the human emotive spectrum, actually are akin to each other. He argues that "for Joubert, the guffaw encompasses the sob, but the reverse is not the case"; Joubert dubs this "a battle of two feelings" (Carroll Simon 268). He draws from Joubert to argue that laughter is "a gift of rest from the body to the mind" (Carroll Simon

268). These protest signs are not always laugh-out-loud funny. The jokes in protest signs are gifts of rest to the minds of people preoccupied with situations of political unrest, such as those disillusioned with the U.K. government's handling of the Brexit referendum. They jest to emphasise the seriousness of the situation. It is an act that subverts authority, as is visible in the double entendre in Fig.18.

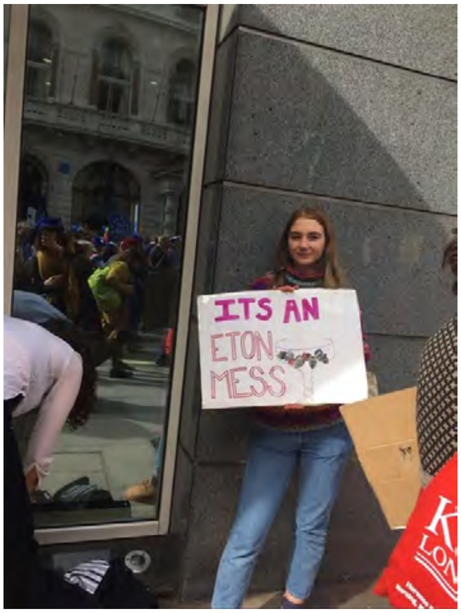


*Figure 18: Klaff, Nina. Eton Mess. Unpublished photograph, London, March 2019.*

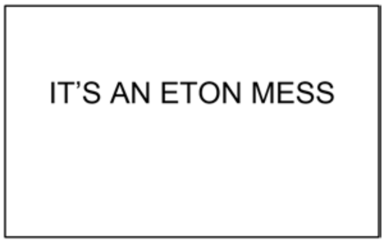


*Figure 19: Klaff, Nina. Illustration of Eton Mess sign at People's Vote March, March 2019.*

Eton Mess refers to a traditional English dessert made of strawberries, meringues, and whipped cream, which is said to have originated at Eton College, the notoriously expensive private school for boys. The school has produced an extraordinary number of British leaders with David Cameron (photographed below, circled in the top row), member of its exclusive Bullingdon Club and the instigator of the Referendum, was the school's 19th alumnus to act as Prime Minister, and Boris Johnson (also photographed below, circled in the bottom row), who became Eton's 20th Prime Minister (Moss).

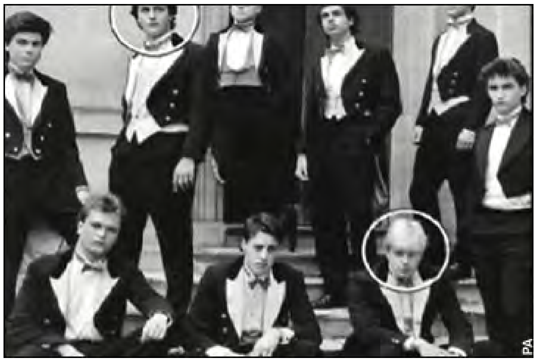


*Figure 20: Unknown Author. "The Bullingdon club of '87." Ben Fenton, The Telegraph, 12 Feb. 2007. Accessed 19/05/2019.*

Eton is considered the pinnacle of the British elite. Eton mess is a symbol of British indulgence that has become part of its gastronomic hegemony. The sign declares the referendum and its consequences to be an Eton mess, a farrago instigated and exacerbated by the leaders who were educated there. Understanding this sign requires an integral understanding of British culture. It also requires a deep understanding of the widespread criticism of the current British government as being led by the elite - and poorly at that. Just as funniness is dependent on shared cultural norms and values, the rhetoric of protest signs often reflects the context in which they were created. The People's Vote March, as an Anglo-specific movement, has a very British character in both its humour and popular culture references, but it has also been influenced by the more global culture of social media aesthetics.

## Social Media Rhetoric and Aesthetics

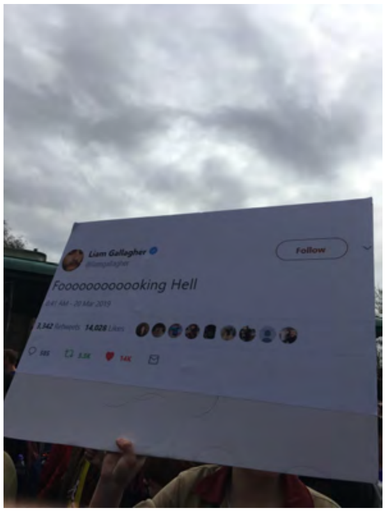


*Figure 21: Klaff, Nina. Tweet by Liam Gallagher as protest sign, unpublished photograph, London, 2019.*

In the image in Fig. 21, the sign is made of a large scale print out of a tweet by British celebrity and former Oasis band member Liam Gallagher glued onto a whiteboard, which reads simply “Foooooooooooking Hell” (@Liam Gallagher). It was posted on the 20th of March 2019, three days before the protest and the date that Prime Minister Theresa May requested a delay in Brexit proceedings from the European Union. Not only are his words included, but so is a screengrab of the original tweet, along with the date, handle, and number of retweets and likes the post received. It remains recognisable as a tweet, exemplifying how, protest signs can be said to have evolved to include social media aesthetics.

Perhaps, a more pertinent example of this is Fig.22, which is a sign that captured my own attention at the People’s Vote March in October 2018.

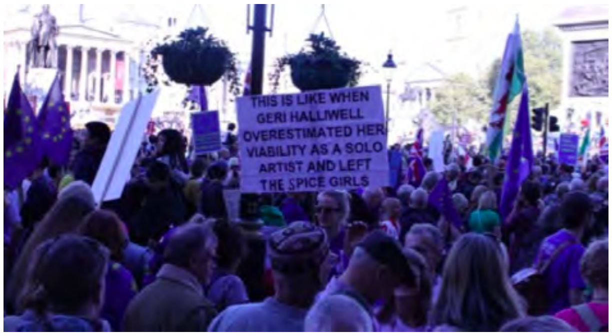
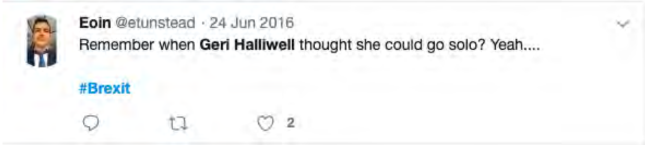


*Figure 22: Klaff, Nina. Halliwell sign, unpublished photograph, London, Oct. 2018.*

THIS IS LIKE WHEN GERI
HALLIWELL
OVERESTIMATED HER
VIABILITY AS A SOLO
ARTIST AND LEFT THE
SPICE GIRLS

*Figure 23: Klaff, Nina. Illustration of Halliwell sign at People’s Vote March, 2019.*

I was drawn to this sign because, quite simply, I personally found it funny. The Spice Girls were idols for those of my generation, children of the 90s, and their split was a British tragedy that resounded on the global stage. The U.K. is here likened to pop singer Geri Halliwell and the E.U. likened to the Spice Girls band, implying that the U.K. has overestimated its strength as a lone country once it is no longer a member of the European Union. Consider the sign’s length: at 102 characters, it is notably longer than usual signs. It is a full sentence, while most signs contain short quips and slogans. Finding the maker of the sign, and quantifying how many times it had been shared on social media, proved difficult: as an image, locating it through social media searches is only possible if it is accompanied by keywords, but there are a number of related tweets of interest. Fig. 24 depicts a Tweet from 2016, mere days after the referendum, comparing Halliwell’s belief in her solo career to the Brexit referendum.



**Eoin** @etunstead · 24 Jun 2016

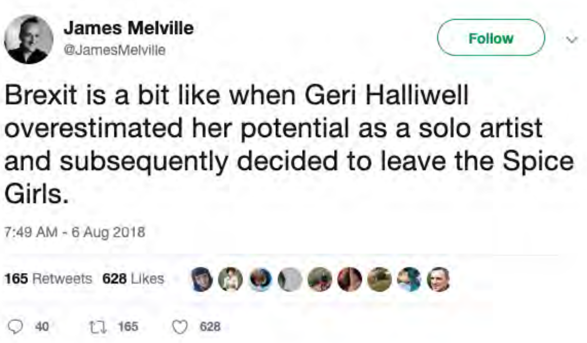
Remember when **Geri Halliwell** thought she could go solo? Yeah....

#Brexit

2

*Figure 24: Screenshot of a tweet by user @etunstead (Eoin). “Remember when. . . ” Twitter, 24 Jun. 2016, 2.47 p.m. Accessed 29/05/2019.*

It received only two likes in response: while this is perhaps an insignificant first iteration of this nature, it was by no means the last. User @James-Melville tweeted, on the 6th of August 2018 (see Fig. 25):



**James Melville**
@JamesMelville

Brexit is a bit like when Geri Halliwell overestimated her potential as a solo artist and subsequently decided to leave the Spice Girls.

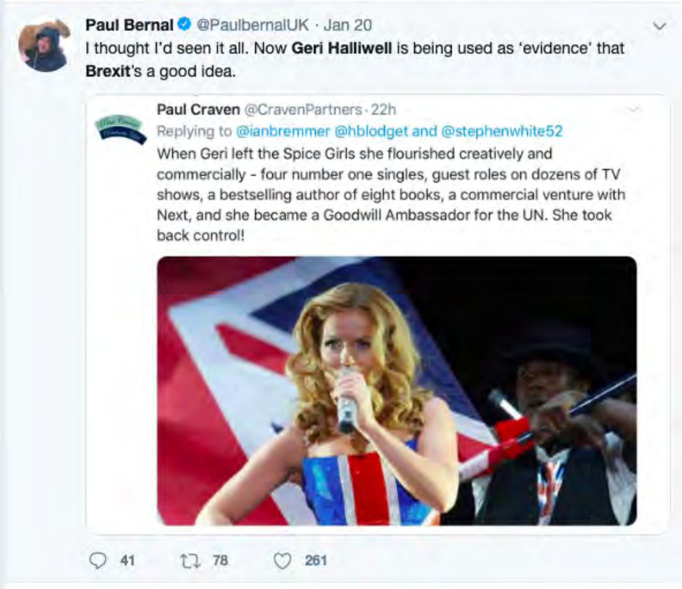
7:49 AM - 6 Aug 2018

**165** Retweets **628** Likes

40 165 628

*Figure 25: Screenshot of a tweet by user @JamesMelville (James Melville). “Brexit is a bit like. . . ” Twitter, 6 Aug. 2018, 7.49 a.m., Accessed 29/05/2019.*

It was retweeted 165 times and received 628 likes. It did not go viral, but it received sufficient shares for it to be at least possible that the author of the sign was overtly or subconsciously inspired by this tweet - or perhaps others. Indeed, the day before the first People’s Vote March, on the 19th of October, Twitter erupted with the debates of this comparison. Halliwell fans also took to Twitter to defend their idol. Some used it as an argument in favour of leaving the EU. @CravenPartners stated that “When Geri left the Spice Girls she flourished” and that through this move “she took back control” (*Twitter*, see Fig.26).



Paul Bernal @PaulbernalUK · Jan 20

I thought I'd seen it all. Now **Geri Halliwell** is being used as 'evidence' that **Brexit**'s a good idea.

Paul Craven @CravenPartners · 22h

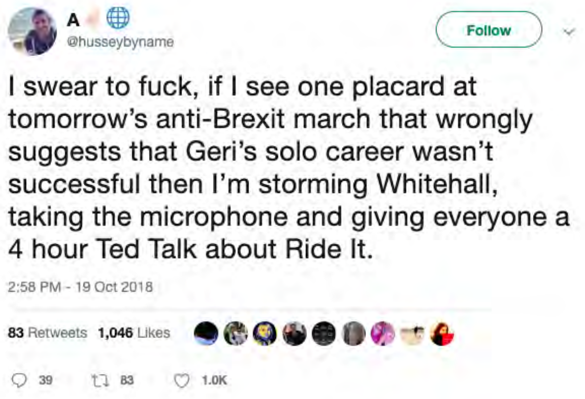
Replying to @ianbremmer @hblodget and @stephenwhite52

When Geri left the Spice Girls she flourished creatively and commercially - four number one singles, guest roles on dozens of TV shows, a bestselling author of eight books, a commercial venture with Next, and she became a Goodwill Ambassador for the UN. She took back control!

41 78 261

*Figure 26: Screenshot of a tweet by user @CravenPartners (Paul Craven). "When Geri left..." Twitter, 19 Oct. 2018, 8.43 p.m. Accessed 20/05/2019.*

Others took the jibe more literally, such as @husseybyname, who swore that if but one sign were to "wrongly suggest that Geri's solo career wasn't successful," they promised to storm the march and lecture on the value of her song "Ride It" (*Twitter*, see Fig.27).



A @husseybyname Follow

I swear to fuck, if I see one placard at tomorrow's anti-Brexit march that wrongly suggests that Geri's solo career wasn't successful then I'm storming Whitehall, taking the microphone and giving everyone a 4 hour Ted Talk about Ride It.

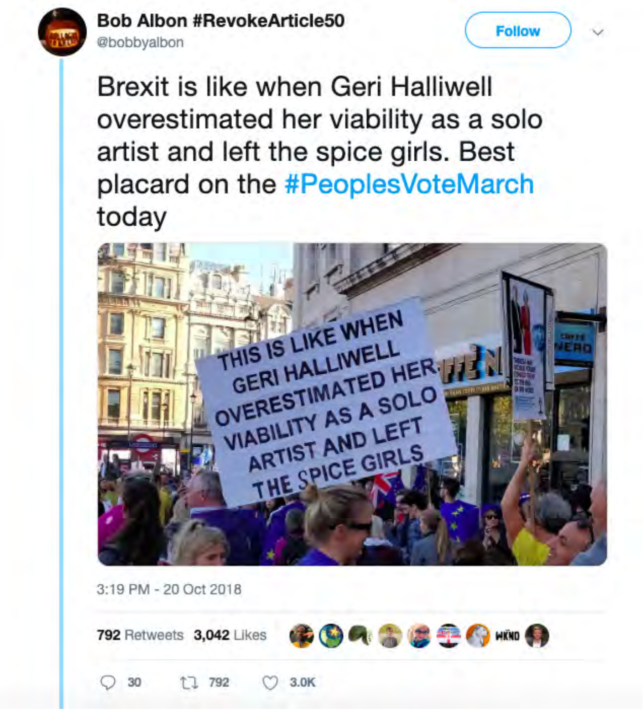
2:58 PM - 19 Oct 2018

83 Retweets 1,046 Likes

39 83 1.0K

*Figure 27: Screenshot of a tweet by user @husseybyname (A). "I swear to fuck..." Twitter, 19 Oct. 2018, 2.58pm. Accessed 20/05/2019.*

It appears whoever made the sign did not take heed of @husseybyname's warning in Fig.27. I was not aware of this context when I took the photo, nor was I the only one drawn to the placard itself. Twitter user @bobbyalbon posted the tweet in Fig.28 while the march was still happening, in real time.



Bob Albon #RevokeArticle50 @bobbyalbon Follow

Brexit is like when Geri Halliwell overestimated her viability as a solo artist and left the spice girls. Best placard on the #PeoplesVoteMarch today

THIS IS LIKE WHEN GERI HALLIWELL OVERESTIMATED HER VIABILITY AS A SOLO ARTIST AND LEFT THE SPICE GIRLS

3:19 PM - 20 Oct 2018

792 Retweets 3,042 Likes

30 792 3.0K

*Figure 28: Screenshot of a tweet by user @bobbyalbon (Bob Albon RevokeArticle50). "Brexit is like when Geri Halliwell. . . " Twitter, 3.19pm, 20/05/2019.* [6]

The sign went on to create a social media storm. Albon's tweet was retweeted over 792 times and received 3,042 likes. *Stylist* Magazine published an article entitled "Forget Brexit, this sign about Geri Halliwell is what's really dividing voters," which was shared all over Facebook (Murray 2018). Humorous engagement of that kind adds to the carnivalesque aspects of protest marches, which has been made possible through social media. The networked public sphere presents a new way of expressing political views through digital subversion and humour that prolongs the life of the protest well after the assembly itself has dispersed.

## III. The Global Women's March

I
CAN'T
KEEP
QUIET

[6] As I observed Fig.29 more closely, attempting to decipher any points of note regarding the image itself, I couldn't help but notice that my mother and I are in the background. She is wearing sunglasses and a blue hat (bottom right), and the stars of my own matching blue E.U. beret are just barely visible behind the person in the yellow shirt with grey hair.

*Figure 29: Klaff, Nina. Illustration: "I can't keep quiet" sign seen at Women's March Amsterdam, 11 Mar. 2017.*

## Context

The first Women's March was held in Washington D.C. on the 21st of January 2017, the day after Donald Trump was inaugurated as the President of the United States. The movement began on social media and its marches and sister marches are still organised online. The movement describes its aims in "Unity Principles": ending violence, fighting for reproductive rights, LGBTQIA rights, workers' rights, civil rights, disability rights, immigrant rights, and environmental justice. They want to "create a global community in which women — including Black women, indigenous women, poor women, immigrant women, disabled women, lesbian, queer, and trans women, and women of every religious, non-religious, and atheist background — are free" (Women's March Global). An analysis of the movement will aim to demonstrate how what was "heralded as the largest single-day synchronized global mass mobilization ever" has continued to mobilise transnationally, and on a global scale, through social media (Tambe 223). I base some of this analysis on my attendance of the 2017 and 2019 "chapters" in Amsterdam, The Netherlands.

## Towards a Networked Public Sphere

Echoing the Arendtian supposition that the space of appearance requires bodily co-presence addressed in Chapter One, Gerbaudo notes that "social movements have traditionally relied on the existence of local face-to-face networks" of embodied co-presence (30). He argues that these "have been considered almost unanimously by scholars as the most important channel for mobilisation" (Gerbaudo 30). However, while both the People's Vote March and Women's Marches culminated in embodied marches, the digital engagement between the participants was, and continues to be, just as important as their embodied manifestation. I argue that social media contributes to, and is inextricable from, these movements' embodied manifestations. As Tufekçi argues, "the whole public sphere, as well as the whole way movements operate, has been reconFigured by digital technologies" (6). Tufekçi proposes the term "digitally networked public sphere" (or "networked public sphere") to explain this new conFiguration of the public sphere (6). She argues that this "does not mean 'online-only' or even 'online-primarily'": It encompasses how the online and face-to-face worlds interact in social movements (Tufekçi 6). Thus, Tufekçi proposes that "the twenty-first-century public sphere is *digitally networked* and includes mass media and public spaces, such as the squares and parks where many protests are held, as well as new digital media" (6, emphasis my own). That is certainly the case in the Women's March, a movement that was catalysed and sustained through digital interaction.

In late 2016, Theresa Shook, a lawyer based in Hawaii, took to social media to voice her outrage at the outcome of the 2016 U.S. presidential election. One significant affordance of digital technologies for social movements is that they give "ordinary people [...] the potential to reach millions of people at once" (Tufekçi 6). According to fellow Women's March organiser Mrinalini Chakraborty, Shook "created a private Facebook event page for the march and invited a few dozen online friends to join before going to sleep" (Tembe 224). As Tufekçi notes, "digital technologies" now "are especially important during the initial formation of social movements" (6; 10). It is understood that Shook first posted the event in a private Facebook group called "Pantsuit Nation," set up for supporters of Hillary Clinton, who ran against Donald Trump in the 2016 election, to communicate with each other (Kairney; Tembe). As Sandoval-Almazan and Gil-Garcia argue, "online social media expands the channels for and velocity of the message," and its benefits are that it incurs "no information costs" and presents "a common ground of technology" (46). Word then spread overnight through Pantsuit Nation and other similar online groups and Shook reported that the movement went viral overnight: "when I woke, up it had gone ballistic," with hundreds of thousands of users declaring their intent to attend the event (Kearney 2016). Then, seeing the viral response, "veteran activists and organizers began planning a large-scale event scheduled for January 21, 2017, the day after Inauguration Day" ("HISTORY" Women's March).

Moreover, not only can digital technologies increase the speed of communication, but they also "alter our sense of space" (Tufekçi 7). Our logic of place has changed (Gerbaudo). What began as a politically charged Facebook post, urging one individual's online friends to mobilise in what was then called the Million Women March in one American city, is now a global movement that operates transnationally, composed of 52 "chapters," or editions, of the movement, with a total of 673 "sister marches" that gathered 4,814,000 marchers in cities around the world ("Women's March" Website). It is now known as the Global Women's March. I argue that the global nature of this movement was only possible because of social media.

However, Tufekçi warns that while social media can help movements spread their message in unprecedented ways, with these affordances also "comes weakness, some of it unexpected" (70). She argues that "the tedious work performed during the pre-internet era served other purposes as well" (xiii). It "acclimatized people to the processes of collective decision making and helped create the resilience all movements need to survive and thrive in the long term" (Tufekçi xiii). According to Chakraboty, "the spontaneous nature of the march's origin meant that not a lot of thought had gone into either the logistical planning of such a mass mobilization or other crucial details, such as the name" (225). As Tufekçi notes, "having arisen so suddenly and grown so quickly," many "networked movements have few means of dealing with the inevitable internal conflicts of politics" (xiii, 270). Indeed, the women's movement came under fire in 2019 as co-president Tamika Mallory reportedly posted a photograph of herself with Louis Farrakhan, the leader of the Nation of Islam, whose extremist views lead to his organisation being declared a "hate group." Vesoulis reports that "in the past, Farrakhan compared Jewish people to termites and described them as 'satanic'". Mallory's post was captioned with "GOAT" - a contemporary slang appraisal, acronym for Greatest of All Time. Mallory's association with a known anti-semite reflected badly on the Women's March organisation. It was vilified in the media and even drew criticism from Theresa Shook herself, who again took to Facebook to argue that this was "in opposition to our Unity Principles" and called for Mallory and others to step down, stating that they had "steered the Movement away from its true course" (Vesoulis). Despite a fall in participation numbers since the controversy, the Women's March continued under the same names, and with the same principles of equality. Continuing to build their community over social media, people around the world, at times with different views, keep taking to the albeit geographically different streets, to enact the same claims of justice. The movement is serialised in that it continues to create new chapters, hold assemblies in new places, and garner support from new participants all over the world. I argue that the Women's Marches' aesthetic formation can be thought of in terms of the ebb and flow of a wave. As Peeren et al. eloquently put it, "while a single wave may fizzle out on the shore, the sea never stops moving" (12). Like a wave, the Women's March movement continues to swell in size, only taking respite until its next iteration, in a different place, becomes active.

## Global Aesthetic Formations

One of the main aims of the Women's March movement is to create a global community of women. In Anderson's *Imagined Communities*, he demonstrates how print media changed our conception of nationhood. He argues that communities such as nations are "*imagined*" as many members will likely never meet, "yet in the minds of each lives the image of their communion" (49). Building on Anderson, Tufekçi demonstrates that "people who would never expect to meet in person or to know each other's name come to think of themselves as part of a group through the shared consumption of mass media" (5-6). She argues that in the 21st century, it is "digital connectivity" that has reconFigured our conception of communities and reshaped how movements connect, organize, and evolve during their lifespan" (Tufekçi xix). In a movement as large as the Women's March, which is dispersed across different continents, most participants will indeed never know many of their fellow-members. Indeed, Meyer argues that Anderson's fleeting remark that "communities are to be distinguished "by the *style* in which they are imagined" (1991, 6, emphasis my own), hints at the importance of scrutinizing how the binding of people into imagined communities actually occurs and is realized in a material sense" (6). However, Meyer asks that we move away from the term "community," as she believes it is limiting: it suggests a "fixed, bounded social group" that is too rigid (6). She posits instead the term "formation," which, she argues, "refers both to a social *entity* (as in social formation) - thus designating a community - and to processes of forming" (Meyer 6-7, emphasis in original). Rather than thinking of these formations as imagined, Meyer argues, we must acknowledge how communities and their imaginations "materialise through media" (7). This necessitates a more explicit focus on things and their thingness itself, but allows us to "[retain] Anderson's view that the making of bonds in modern times depends on media and mediation" (Meyer 7). Thus, a consideration of the Women's March, a movement that was organised, dissipated, and sustained through social media on a global scale, requires a consideration of the influence of social media and digital practices in binding people within the movement. Meyer argues, "more attention needs to be paid to the role played by things, media, and the body in actual processes of community making" (6). She posits the term "aesthetic formations" to argue that "community thus evolves around shared images and other mediated cultural forms" (9). I argue that the aesthetic formation of the Women's Marches developed through social media, and that many of the shared images that constructed their aesthetic formations were protest signs that were shared online.

## Social Media, Global Culture

Much of the aesthetic visible in the signs carried by protesters was evidently informed and influenced by social media rhetoric, literacy, and aesthetics. In the sign below, an activist holds up a piece of cardboard inscribed with the words "JUST, UGH." in capital letters (Fig.30).

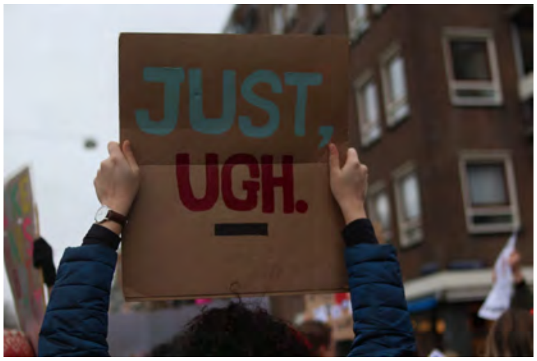


*Figure 30: Klaff, Nina. "JUST, UGH." Unpublished photograph, Women's March, Amsterdam, 9 Mar. 2019.*

This is a turn of phrase commonly used on the internet to express exasperation. The onomatopoeic "UGH" is a sound that hits the back of the throat, an expression of disgust, frustration, anger even. The expression here represents a universal disillusionment with the status quo. The piece of found cardboard, crumpled at the bottom and with a line down the middle where it has been bent in two, is held up defiantly by an activist, acting as an extension of their Figure. The efforts to make the words look like printed typing demonstrate the crossover of social media and the internet aesthetic in the grassroots ethos of the movement. Thus, the Women's Marches' protest signs, just as those of the People's Vote march, have also become influenced by contemporary social media aesthetics.

Another key way in which this is visible is in the similarity of the signs visible in the chapters in entirely different locations. Consider the sign below, held by an activist at the Amsterdam chapter in March 2019 (Fig.31)

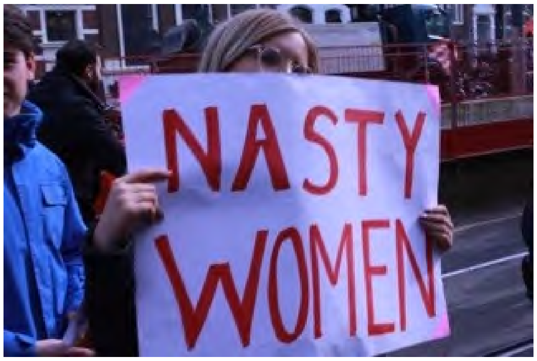


*Figure 31: Klaff, Nina. "NASTY WOMEN." Unpublished photograph, Women's March, Amsterdam, 11 Mar. 2017.*

The phrase refers to an instance in which Trump called his opponent Hillary Clinton a "nasty woman" during one of the final presidential debates. The Guardian reports that "NastyWoman was soon trending as people reacted against Trump" (Woolf). It became a slogan of the movement. It was turned into a poem by Nina Mariah Donovan and performed emphatically by Ashley Judd on the 21st January 2017, the video receiving over a million views (*YouTube*). The term also appeared on a protest sign in Amsterdam, and emerged in different in marches in other cities - in other countries.

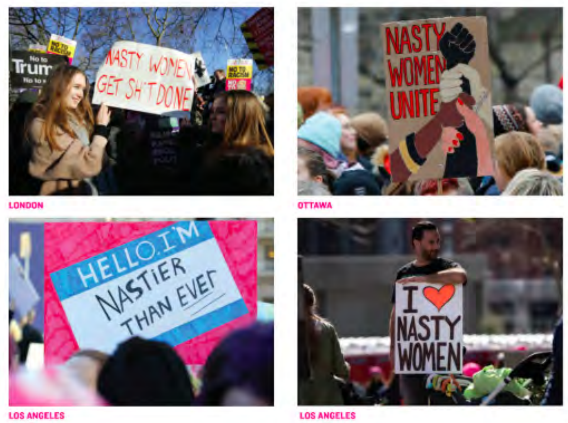


*Figure 32: Artisan. Scan of Page 22 of Why We March: Signs of Protest and Hope - Voices from the Women's March. Artisan, New York, 2017.*

In *Why We March: Signs of Protest and Hope - Voices from the Women's March*, an entire book devoted to the protest signs of the Women's Marches all over the world, the signs that refer to the "Nasty Woman" debacle are given a whole page spread of their own. They are here documented as appearing in London, Ottawa, and L.A. This demonstrates that the rhetoric of protest signs is influenced by trending and viral events in mainstream and social media and the humour is subversive. The images and slogans its participants produce have been shared and spread widely and across distances, countries and continents. Thus, just as the movement claims globality, so too might its aesthetic formation.

## Epilogue: Online Protest Signs

*Figure 33: Klaff, Nina. Illustration of "accessibility sets us free" sign, as seen at Women's March Amsterdam, 9 Mar. 2019.*

As their name implies, marches entail walking: through public space, in protest, together. The ability to do so is often taken for granted. If marches continue to be the epitome of politics, the space of appearance continues to be thought of only in terms of bodily co-presence: this excludes those not able to access these spaces. The Arendtian conception of the space of appearance as contingent on embodied co-presence public sphere does not account for those who are not able to be present for whatever reasons. Indeed, in *Sick Woman Theory*, Hedva takes issue with Arendt's presupposition that political activism requires bodily presence in the public sphere: she says, "Arendt said, just get your body into the street, and bam: political" (Hedva). However, as Hedva demonstrates,

> The inevitability of violence at a demonstration – especially a demonstration that emerged to insist upon the importance of bodies who've been violently un-cared for – ensures that a certain amount of people won't, because they can't, show up. Couple this with physical and mental illnesses and disabilities that keep people in bed and at home, and we must contend with the fact that many whom these protests are for, are not able to participate in them – which means they are not able to be visible as political activists. ("Sick Woman Theory")

As Hedva pertinently asks, "how do you throw a brick through the window of a bank if you can't get out of bed?" (Hedva). If political activism requires embodied participation, how can one participate if one's body is not safe in the public sphere, or if the space of appearance itself is not accessible? Indeed, how can one hold up a protest sign, if one can't get out of bed? The answer may well lie in digital media once again.

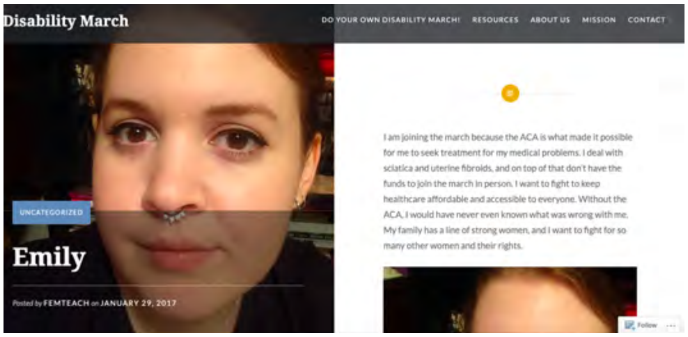
In the Amsterdam chapter of the Women's March, the marching assembly convened on Museumplein, where protesters gathered to hear speeches by organisers and activists. The second speaker was Annika Mell, a disability activist who pertinently explained: "I didn't march with you today" and stressed that this was "not because I don't care" (Mell). She told the crowd how her physical disability prevented her from participating in the march, and that her experience is not an isolated case:

> "*We*" didn't march with you today because *I* am not alone in this. *We* didn't march with you today because we are too sick, too tired, too fragile, in too much pain, too anxious, too depressed, too afraid of big cities, too afraid of big crowds to participate in a massive event like this. We didn't march with you today because by its very nature marches such as this are not accessible to everyone. [...] We

didn't march with you today because we are sick and tired of being excluded from feminist spaces. (Mell 2019, transcript and emphases my own)

Mell's vocal condemnation of the persistent inaccessibility of not only the Women's March but the feminist movement in general demonstrates an inconsistency in the claims of equality and intersectionality. The Women's March movement have made efforts to address this through holding online Disability Marches.

The Washington Disability March began on December 21st 2018. According to their website, by 7pm on 21st of January, they had "1,654 entries published, and thousands more" that were not processed (Disability March). That number represents those who could not, for whatever reason, attend the embodied march, but wished to remain visible in the space of appearance. On this page, participants posted representations of themselves that acted as signs of their participation. They uploaded images of themselves. Recall here Sontag's claim that "a photograph is both a 'pseudo-presence' and a 'token of absence'" (qtd. in Villi 2). Building on this, Villy argues, "photographs offer presence-in-absence" (Villi 3). Moreover, these "tokens of absence" on the DM are not only visual, but also textual. I argue that these would be the Disability March's equivalent to the protest sign. While they did not participate in the embodied marches, their presence within the movement is enacted through their online postings. While the protest placard is traditionally limited to short, memorable slogans, the posts on both the Washington and Amsterdam sites go into great detail. They describe their reasons for absence, which are often extremely personal. They explain what has kept them from marching (e.g. physical or mental illness, absence from country), and why they want to participate in the movement. "Emily" posted, "I deal with sciatica and uterine fibroids, and on top of that don't have the funds to join the march in person. I want to Fight to keep healthcare affordable and accessible to everyone" (Fig. 34).



Emily

Posted by FEMTEACH on JANUARY 29, 2017

I am joining the march because the ACA is what made it possible for me to seek treatment for my medical problems. I deal with sciatica and uterine fibroids, and on top of that don't have the funds to join the march in person. I want to fight to keep healthcare affordable and accessible to everyone. Without the ACA, I would have never even known what was wrong with me. My family has a line of strong women, and I want to fight for so many other women and their rights.

*Figure 34: Screenshot of "Emily's" participant contributions to the Disability March. Disability March, 21 Jan. 2019.*

While she was not able to join the march in person, her voice is not to be discounted: here, on the Disability March, she can vocalise her desire "to Fight for so many other women and their rights" (Fig.34). As Tufekçi argues, the "availability" of media "not only affects the kinds of discussions that can be had, [but also] the kinds of people who can have them" (7). Media can thus be positive and inclusionary forces within movements. The standalone website devoted to those not able to march gave platform for their expression of their participation in a movement, and I thus argue that these posts are to be thought of in equal standing as those displayed at the embodied assemblies. Moreover, online assembled collectivities remain long after the gathering of an assembly has dispersed as their digital footprints remain online, and they hold the potential to be re-engaged with time and time again.

With this in mind, I argue that there are other objects that might be considered as protest signs that were not covered in the limited scope of this paper. The People's Vote March, for example, created a petition to "Revoke Article 50 and Remain in the E.U.," which, despite receiving over 6 million digital signatures, was unsuccessful in its aim (Geordianou).[7] The government's response was again to assert its promise "to honour the result of the 2016 referendum" (Geordianou). However, each of those signatures acts as an autonomous sign of dissent. While these digital signatures do not encompass the embodied, expressive, and aesthetic aspects of protest signs, they are of equal standing as their markers of participation within a movement; they are also persuasive acts. Individuals may use visual media in their online identity

[7] The government made this statement in response: "This Government will not revoke Article 50. We will honour the result of the 2016 referendum and work with Parliament to deliver a deal that ensures we leave the European Union" (Geordianou).

to perform protest digitally, by using filters on their Facebook profile pictures for example.

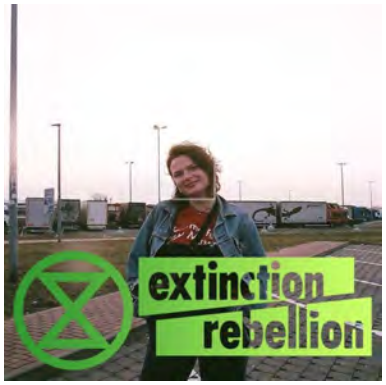
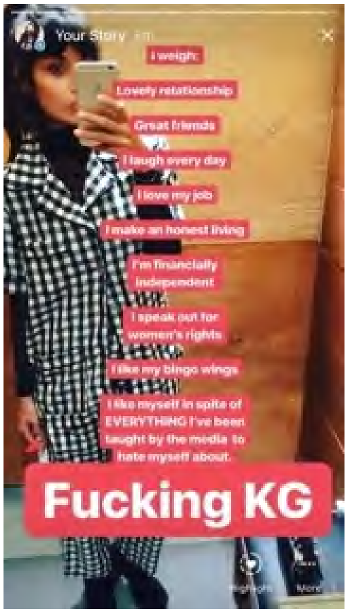


*Figure 35: User Kate Mitchell's profile picture with climate movement Extinction Rebellion's filter, Facebook, 12 Feb. 2019. Accessed 30/05/2019. Included with permission from user.*

Moreover, there are also many, many ways of using social media to build collectivities that are not aimed to be conducive to culminating in embodied assemblies, but are created only to operate online. One such example is the selfie-feminism @iWeigh movement, founded by celebrity Jameela Jamil, who posted the image below on her Instagram story out of outrage at the societal obsession with women's weight.



Your Story

I weigh:

Lovely relationship

Great friends

I laugh every day

I love my job

I make an honest living

I'm financially independent

I speak out for women's rights

I like my bingo wings

I like myself in spite of EVERYTHING I've been taught by the media to hate myself about.

Fucking KG

*Figure 36: Screenshot from @iWeigh instagram page: Jamil, Jameela. "I Weigh. . . Fucking KG." Instagram, 2018. Accessed 29/05/2019.*

Jamil posted an image of herself with all the other things she believes to be more important than the numbers on a scale, such as her "lovely relationship," "great friends," and her liking herself "in spite of EVERYTHING" she was "taught by the media to hate" (@iWeigh). Through her already vast online platform, this image instigated a movement of more than 742,000 followers who participate by sharing similar photos of themselves. I argue that these are also to be considered aesthetic and embodied expressions of dissent that contribute to the creation of aesthetic formations of movements. While this project has focused on material and embodied protest signs, the digitally networked nature of the 21st century public sphere makes such limited consideration insufficient. I therefore encourage further research into the digital mutations of embodied protest signs that goes beyond sharing images of signs on social media.

## Conclusion: What do Protest Signs Want?

*Figure 37: Klaff, Nina. Illustration paraphrase of George Orwell, "Every Joke is a Tiny Revolution." 2019.*

Social movements arise when people believe that "a wrong should be righted or a change should be made" (Tufekçi 25). In the two case studies I have addressed, firstly, the British people gathered in London to demand that the Brexit wrongs be righted, and secondly, at Women's Marches all over the world, people have been demanding equal rights for women everywhere. I reiterate that while these were peaceful demonstrations, both movements arose out of a perceived threat to the liberty and safety of a people, and the stakes remain high for both. I have not aimed to ascertain whether or not these movements have been successful. This is because I do not believe that the success of protest movements can merely be assessed on "instrumental aims" (Tufekçi 89). While the British government remains resolute in proceeding with Brexit, Donald Trump is still in office, and the patriarchal system that the Women's Marches so fervently aim to undermine is still very much in operation, as Tufekçi argues, "protests can be ends as well as means. People wish to belong, especially to communities that make them feel good" (Tufekçi 90). I argue that protest marches, as performative rituals - carnivalesque spectacles of popular dissent - can be events in which people come together, united in a singular aim, and that this in itself can be a positive force for good. These "festivals of resistance" are indeed carnivals "in which the individual life celebrates its unification with a regenerated society" in contemporary protest culture (Graeber 23; Vaneigem qtd. in Tancons 297). I argue that the creation and display of protest signs are an important way of making a people's unification and their desire for a regenerated society explicit.

Indeed, as Rieff argues, "art is a means of union among [people], joining them together in the same feelings," which is "indispensable for their community, however that community is conceived" (479). The artistic production and emphatic display of protest signs unites people in the protest event. They are thus to be understood as forms of activist art, or artivism, truly indispensable in the creation of "aesthetic formations" of social movements (Meyer 9). They are invaluable props to the dramaturgy of protest events through which individuals express their political thoughts and emotions. They are integral objects to the "repertoires of contention" of many social movements that allow for participants to display their autonomy through creative originality, even as they become one with a crowd and swept up in "collective effervescence" (Tufekçi 89; Durkheim). They are also important objects of cultural production that reflect the norms and values of the context in which they are displayed: the politicised, and notably British, humour in the signs at the People's Vote March, and the prevalence of social media aesthetics and rhetorical influences in the signs both in London and at Women's Marches around the world, display their nature as artifacts of "participatory culture," that reflects the context of the movement and the shared norms and values of its participants (Jenkins). Moreover, because they are standard features in contemporary protests, the signs are easily recognisable objects of dissent that often become the subject of attention themselves: they might go viral on social media, or be used as headline photos for news reports. The "best signs" at both the People's Vote and Women's Marches marches were featured in articles in news multiple news outlets, from *The Guardian* to *Vogue*, and *Artisan* even published a complete anthology of the Women's March signs, called *Why We March: Signs of Protest and Hope - Voices from the Women's March*, which demonstrates their significance as emblems of dissent (2018).

The messages they carry continue to resonate as they are published and shared on the global media stages, even as their function as embodied protest placards might vanish with the dispersal of the assembly, and become akin to the banner or poster.

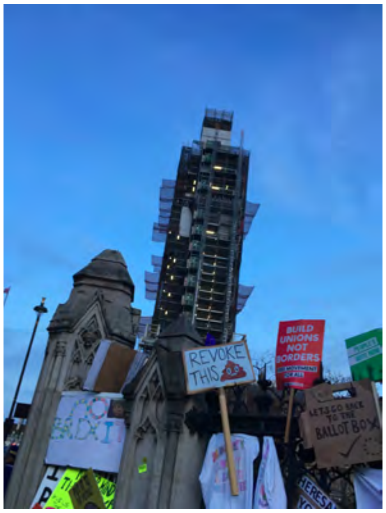


*Figure 38:* Klaff, Nina. Protest signs stuck to the gates of the Houses of Parliament after the People's Vote March, unpublished photograph, 23 Mar. 2019.

The Arendtian supposition of the "space of appearance" being ephemeral and contingent on the embodied co-presence in one place no longer holds true (*The Human Condition*). Indeed, as Tufekçi argues, the 21st century public sphere is "digitally networked," as mainstream and social media enable its extension in both time and space (6).

Moreover, while the two movements in questions might not have achieved their explicit aims, there is another marker of their success. I agree with Tufekçi that a key indicator of the strength of a movement lies in its "narrative capacity," how it might "frame its story on its own terms" and "spread its worldview" (Tufekçi 192). I argue that protest signs are invaluable devices in constructing this narrative capacity, and that protest signs should be the subject of more research as material and cultural objects of socio-political import within protest discourse. I propose that further research might include the signs produced and circulated online-only.

People march because they care. They use protest signs to demonstrate why and in what way they care. Aesthetic devices, under Rieff's definition, communicate feelings, in the hope that those to whom the messages are transmitted "also experience them" (479). When Mitchell asks "*what do pictures really want*?" he accepts that the "answer may well be, nothing at all" (81). However, protest signs are objects of artistic expression, as well as expressions of aesthetics, rhetoric, and persuasion; what they want is never nothing at all. They want for the people to be asked what they want and for this to be heard. When people march, they are making demands. They express their demands in the signs they make and hold themselves, and they must be given attention. Whether they are humorous or just plain angry, to paraphrase Orwell: every *sign* is a tiny revolution.

## Works Cited

"About." *PussyHat Project: Design Innovations for Social Change*.

Adams, Bruce. *Tiny Revolutions in Russia: Twentieth-century Soviet and Russian history in anecdotes*, Routledge Curzon, 2005.

Anderson, Benedict. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, London: Verso, 1983.

Arendt, Hannah. *The Human Condition*. 2nd ed., University of Chicago Press, Chicago, 1958.

Artisan. *Why We March: Signs of Protest and Hope - Voices from the Women's March*. Artisan, New York, 2017.

BBC News. "People's Vote march: Hundreds of thousands attend London protest." *BBC News* Online, 20/10/2018.

BBC News. "EU Referendum: Results in Full." *BBC News Online*, 2019. Accessed 7/05/2019.

Berlant, Lauren and Sianne Ngai. "Comedy Has Issues." *Critical Inquiry 43*, no. 2, Winter 2017, pp. 233-249. https://doi.org/10.1086/689666

@bobbyalbon (Bob Albon RevokeArticle50). "Brexit is like when Geri Halliwell overestimated her viability as a solo artist and left the spice girls. Best placard on the PeoplesVoteMarch today." *Twitter*, 3.19pm, 20/05/2019.

Bowcott, Owen. "'Corrupt' Vote Leave campaign undermines Brexit vote, court told." *The Guardian Online*, 7 December 2018. Accessed 5/05/2019.

Butler, Judith. *Notes Toward a Performative Theory of Assembly*. Harvard University Press, Cambridge, MA, 2015, 154-92.

Carroll Simon, David. "The Anatomy of Schadenfreude; or, Montaigne's Laughter." *Critical Inquiry 43*, no. 2, Winter 2017, pp. 250-280.

"Charlie Chaplin - Protest Scene - Modern Times 1936." *YouTube*, uploaded by Charlie Chaplin Official, 4th February 2019, Retrieved 1/3/2019.

Couldry, Nick, et al. "Collectivities." *The Mediated Construction of Reality*, 2017.

Disability March, *Disability March*, 21 Jan. 2019. Accessed 19/05/2019.

Downey, Anthony. *Art and Politics Now*. Thames Hudson, London, 2014.

Durkheim, Emile. *The Elementary Forms of Religious Life*. translated by Karen E. Fields, The Free Press, New York, 1995. mbaker/cshs503/durkheimreligiouslife.pdf.style

Emily. "Why I Am Joining the March." Disability March, Posted by FEMTEACH on 29/01/2019. Accessed 5/06/2019/.

@etunstead (Eoin). "Remember when Geri Halliwell thought she could go solo? Yeah. Brexit." *Twitter*, 24 Jun. 2016, 2.47 p.m. Accessed 29/05/2019.

Faulconbridge, Guy, and Kate Holton. "'Explosive shock' as Britain votes to leave EU, Cameron quits." *Reuters*, 23rd June 2016. Accessed 8/05/2019.

"'Fromage not Farage': the best signs and sights on the People's Vote march'." *The Guardian*, 23 March 2019. Accessed 7/05/2019.

Garlough, Christine. "Vernacular Culture and Grassroots Activism: Non-Violent Protest and Progressive Ethos at the 2011 Wisconsin Labour Rallies." *The Political Aesthetics of Global Protest: the Arab Spring and Beyond*, edited by Pnina Werbner et al., Edinburgh: The Edinburgh University Press, 2014,

Gerbaudo, Paolo. *Tweets and the Streets: Social Media and Contemporary Activism*, Pluto Press, London, 2012, pp. 18–47. *JSTOR*. Accessed 21/2/2019.

Geordianou, Margaret Anne. “Revoke Article 50 and Remain in the E.U.” *Petitions*, UK Government and Petitions, 20 Aug. 2019. Accessed 31/05/2019.

Holm, Nicholas. *Humour as Politics: the Political Aesthetics of Contemporary Comedy.* Palgrave Studies in Comedy, 2017.

@JamesMelville (James Melville). “Brexit is a bit like when Geri Halliwell overestimated her potential as a solo artist and subsequently decided to leave the Spice Girls.” *Twitter*, 6 Aug. 2018, 7.49 a.m. Accessed 29/05/2019.

Jamil, Jameela. @iWeigh. *Instagram Account*

Jenkins, Henry, et al. *Confronting the Challenges of Participatory Culture: Media education for the 21st century*. MIT Press, 2009.

Kapferer, Judith. “Constructing a Public Sphere: Materiality and Ideology.” *Social Analysis: The International Journal of Social and Cultural Practice*, vol. 51, no. 1, 2007, pp. 68–85. JSTOR,

Kearney, Laila. “Hawaii grandma’s plea launches women’s march in Washington.” *Reuters*,

Kentish, Benjamin. “Brexit: Campaign to prosecute Boris Johnson over £350m NHS bus claim raises £24,000 in two days.” *The Independent*, 8 Sept. 2018. Accessed 29/05/2019.

McQuiston, Liz. *Visual Impact: Creative Dissent in the 21st Century*. Phaidon, London, 2015.

Mell, Annika. Speech at Women’s March Amsterdam, 10th March 2019. Footage available on YouTube. Women’s March on Amsterdam 2019 | Annika Mell speech | March 9 2019, Uploaded by Women’s March, The Netherlands, published 14th April 2019. Accessed 6/05/2019.

Meyer, Birgit. *Aesthetic Formations*. Palgrave Macmillan, 2009.

Mitchell, W. J. T. “Image, Space, Revolution: The Arts of Occupation.” *Critical Inquiry 39.1*, 2012, 8-32.

Mitchell, W. J. T. “What Do Pictures ‘Really’ Want?” *Octob*er, vol. 77, 1996, pp. 71–82. *JSTOR*,

Moss, Paul. “Why has Eton produced so many prime ministers?” *BBC News*, 12/05/2010.

Montaigne. “Comme nous pleurons et rions d’une même chose.” *The Complete Essays of Montaigne*.

Murray, Megan. “Forget Brexit, this sign about Geri Halliwell is what’s really dividing voters.” *Stylist*, 2018.

Peeren, Esther et al. *Global Cultures of Contestation*. Springer International Publishing, 2018.

Philipps, Axel. “Visual Protest Material as Empirical Data.” *Visual Communication*, vol. 11, no. 1, 2012, pp. 3–21. https://doi-org.proxy.uba.uva.nl:2443/10.1177/1470357211424

Rabinovich, Ben. “Brexit: Everything you need to know about the People’s Vote UK march,” *Mail Online*, 19th October 2018. Accessed 14/03/2019.

Rancie`re, Jacques. *Dissensus: On Politics and Aesthetics*, ed. and trans. Steven Corcoran, Continuum, 2010.

Reflect. “Ashley Judd’s EPIC "Nasty Woman" Speech At The Women’s March On Washington.” *YouTube*, 21 Jan. 2017. Accessed on 21/05/2019.

Reiss, Matthias. ed. *The Street as Stage: Protest Marches and Public Rallies since the Nineteenth Century*. Oxford University Press, 2007.

Rieff, Philip. “Aesthetic Functions in Modern Politics.” *World Politics*, vol. 5, no. 4, 1953, pp. 478–502. JSTOR,

Serafini, Paula. “Subversion through Performance: Performance Activism in London.” *The Political Aesthetics of Global Protest: the Arab Spring and Beyond*, edited by Pnina Werbner et al., Edinburgh: The Edinburgh University Press, 2014.

SharpEdges. “Women’s March Amsterdam 2019. De speech van Naomie Pieter (Naomi Balentien).” *YouTube*, 9 Mar. 2019. Retrieved 5/5/2019.

“Sister Marches.” Women’s March On Washington. Archived from the original on January 23, 2017. Accessed 7/5/2019.

Slackman, Michael. “When a Punch Line is No Longer a Lifeline for Egyptians.” *New York Times*, 5 April 2011. Accessed 28/05/2019.

Tambe, Ashwini. “The Women’s March on Washington: Words from an Organizer: An Interview with Mrinalini Chakraborty.” *Feminist Studies*, vol. 43, no. 1, 2017, pp. 223–229. *JSTOR*,

Tancons, Claire. “Occupy Wall Street: Carnival Against Capital? Carnivalesque as Protest Sensibility.” *The Political Aesthetics* *of Global Protest: the Arab Spring and Beyond*, edited by Pnina Werbner et al., Edinburgh: The Edinburgh University Press, 2014.

Taussig, Michael. “I’m so Angry I Made a Sign.” *Critical Inquiry*, vol. 39, no. 1, 2012, pp. 56–88. *JSTOR*,

Trentmann, Frank. “Materiality in the Future of History: Things, Practices, and Politics.” *Journal of British Studies*, vol. 48, no. 2, 2009, pp. 283–307. *JSTOR*,

Tufekçi, Zeynep. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, New Haven London, 2017.

Vesoulis, Abby. “Women First Marched to Challenge Trump. Now They Are Challenging Each Other.” *Time Magazine*, 19/01/2019. Accessed 11/05/2019.

Werbner, Pnina, et al. *The Political Aesthetics of Global Protest: the Arab Spring and Beyond*. Edinburgh: The Edinburgh University Press, 2014.

2